Association for Information Systems
AIS Electronic Library (AISeL)

MWAIS 2019 Proceedings | Midwest (MWAIS)

5-21-2019

Decision Support for Data Virtualization based on Fifteen Critical Success Factors: A Methodology

Marwin Shraideh
Technische Universität München, marwin.shraideh@in.tum.de

Matthias Gottlieb
Technische Universität München, matthias.gottlieb@tum.de

Isabel uhrmann
Technische Universität München, i.fuhrmann@tum.de

Harald Kienegger
Technische Universität München, harald.kienegger@in.tum.de

Markus Böhm
Technische Universität München, markus.boehm@in.tum.de

See next page for additional authors

Authors
Marwin Shraideh, Matthias Gottlieb, Isabel uhrmann, Harald Kienegger, Markus Böhm, and Helmut Krcmar

Decision Support for Data Virtualization based on Fifteen Critical Success Factors: A Methodology

Marwin Shraideh
Technical University of Munich
marwin.shraideh@in.tum.de
Matthias Gottlieb
Technical University of Munich
matthias.gottlieb@tum.de
Isabel Fuhrmann
Technical University of Munich
i.fuhrmann@tum.de
Harald Kienegger
Technical University of Munich
harald.kienegger@in.tum.de
Markus Böhm
Technical University of Munich
markus.boehm@in.tum.de
Helmut Krcmar
Technical University of Munich
krcmar@in.tum.de

ABSTRACT

Data analysis is important for creating a competitive advantage, but the amount of data is already massive and increasing rapidly. Practitioners are looking for general models for different use cases in deciding whether to virtualize data or not and when it is applicable. However, there is a research gap in such models. Thus, in this study, we applied a design science approach in a further step to develop an IT artifact. It is derived from 15 critical success factors, building the foundation for a heuristic individual decision support on data virtualization. In addition, we calculate a final score that recommends extract transfer and load (ETL), hybrid, or data virtualization. The score adapts flexibly to business needs and helps practitioners make decisions. This IT artifact extends the knowledge base by a new methodology for decision support in data virtualization.

Keywords

Data virtualization, Heuristic, Adaptive data virtualization score, Decision support.

INTRODUCTION

In 2025, the estimated yearly volume of data generated worldwide will have quintupled (Statista, 2018). Therefore, the strategic importance of analyzing this data to support informed decisions is increasing since it has the potential to create a competitive advantage (LaValle, Lesser, Shockley, Hopkins and Kruschwitz, 2011; Rao, McNaughton and Mansingh, 2018; Schroeck, Shockley, Smart, Romero-Morales and Tufano, 2012). However, to analyze this huge amount of distributed data, a flexible approach for connecting to "*heterogeneous, non-conventional internal[,] and external sources of data*" (Rao et al., 2018) is needed. This expedites experimenting, prototyping, and evaluating various analytic initiatives, thus providing better control of the implementation's scope, cost, and timeline. (Rao et al., 2018) In this context, "*data virtualization (DV)*" (Van der Lans, 2012) plays a crucial role compared to traditional extract, transfer, and load (ETL)/data warehousing (Rao et al., 2018). DV *"is the technology that offers data consumers a unified, abstracted, and encapsulated view for querying and manipulating data stored in a heterogeneous set of data stores.*" (Van der Lans 2012, p. 4) It can be used for many applications, such as real-time business intelligence (BI), enterprise-wide search, or scalable transaction processing; thus, it enables exposure to big-data analytics, integrating views across multiple domains with high performance and improvement in security and access, to only name a few (Yuhanna and Gilpin, 2012, p. 3). Besides these features and advantages, many factors influence the success of using or implementing DV for data integration, making ETL or a hybrid approach with ETL and DV necessary. These factors are further called critical success factors (CSF). Currently, CSFs have been identified (see, e.g., Gottlieb et al., 2019). However, they are only observed in various combinations for specific use cases. To the best of our knowledge, a decision support system recommending DV or ETL for different use cases and considering these CSFs has not been developed yet.

However, this paper addresses these issues by designing and implementing an IT artifact. Thus, we develop a decision support heuristic that extends the knowledge base. We consider 15 CSFs to advise to what extent and when DV should be preferred. This methodology provides structured and systematic decision support for practitioners to decide upon ETL, DV, or a hybrid solution with ETL and DV as a suitable data integration approach.

METHOD

In a previous step, we identified 15 CSFs (Gottlieb et al., 2019). We supplement these findings with the design and implementation of an IT artifact, according to (Hevner, 2007), as shown in Figure 1. In this article, we focus on the technical applicability and the targeted functionalities based on high-level criteria. Therefore, the 15 CSFs need to be rated and weighted by the user. Statements defining a unique characteristic, or an example, of each CSF support the rating. The users rate the extent to which a statement is met on a seven-point, bipolar Likert-Scale, which ranges from one to seven to increase assessment variance and receive a valuable assessment of neutral characteristics due to an uneven scale (Cox, 1980). One represents a low match of a CSF and ETL as a recommended approach. Seven represents a high match and recommends DV accordingly. To identify participants who are not reflecting their answers, we also included inverted cases (Bühner, 2012). The goal is to achieve a differentiated rating and an appropriate data distribution. For weighting each CSF, users must assign a score between one and five, where one reflects the lowest level of importance and five the highest. A weight of zero excludes a specific CSF. This weighting is mandatory since it is impossible to provide an overall weight that applies to all data integration problems due to unique requirements, priorities, and goals of each data integration project, as well as different stakeholders' interests.

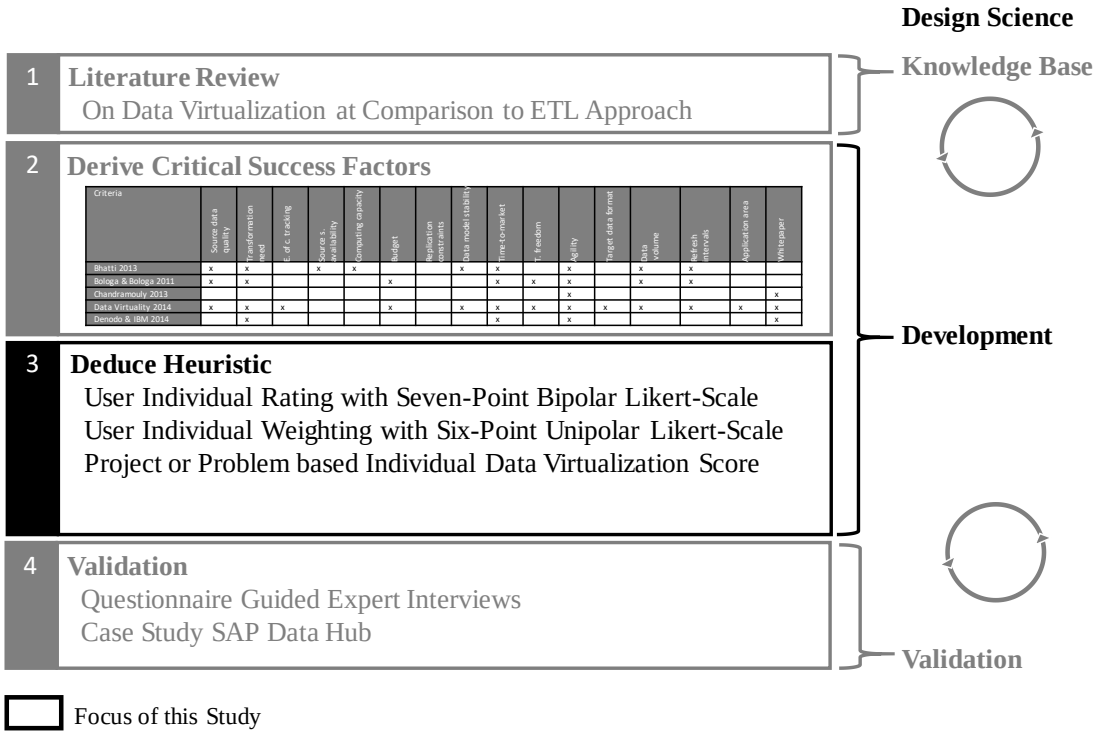

Figure 1. Research Methodology, According to Gottlieb et al. (2019); Hevner, March, Park and Ram (2004)

RESULTS

In a prior study, we identified 15 CSF (Gottlieb et al., 2019). Table 1 presents definition for every CSF.

CSF	Definition
Source data quality	determines the estimated amount of effort needed for data cleansing steps, such as removing duplicates or incomplete tuples (Denodo, 2014). Bad data quality, such as redundant data, favors choosing a physical consolidation approach (Bhatti, 2013).
Transformation need	is the transformation activities needed to integrate data from a source schema into a target schema. Multiple transformation steps decrease DV performance drastically (Bologa and Bologa, 2011; Denodo, 2014).
Extent of Change-Tracking	describes to which extent changes in the source system need to be tracked. (Van der Lans, 2012). For tracking changes with a great extent, physical replication is required (Moxon, 2015; Vinay, 2012).
Source system availability	is the overall stability and reliability of a source system. It is one of the main requirements to be able to utilize DV (Van der Lans, 2012).
Computing capacity	is the remaining computing power of a source system that can be utilized without performance losses. A balanced source system utilization is a significant criterion for the effective and efficient implementation of DV (Van der Lans, 2012).
Budget	is the cost framework for the project, which defines the limits of possible actions such as developing data integration solutions (Voet, 2018). Therefore, it influences the decision on DV.
Replication constraints	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, DV is the approach of choice (Van der Lans, 2012).
Data model stability	describes how often changes in the data model of a source system are made (Marche, 1993).
Time-to-market	is *"the amount of time it takes to design and manufacture a product before it is available to buy"* (Cambridge University Press, 2019). In our case, it is the time taken to design and implement a data integration solution.
Technology freedom	describes the required flexibility to independently choose from many solutions of different vendors (Shankar, 2017).
Agility	is the possibility to react on changes in fast-paced business environments (Bhatti, 2013) with adapting the structure of underlying source systems.
Target data format	is the necessary availability of needed data structure in the source system. The efficiency of data integration depends on the chosen data format (Denodo, 2014, p. 5).
Data volume	is the amount of accessed data. DV enables reading and transforming data on demand and processes it while reading (Bologa et al., 2011).
Refresh intervals	is the frequency of data updates in the source system (Farooq, 2013).
Application area	describes the analytical workload necessary to get the expected results, such as data mining or predictions (Denodo, 2014, p. 5).

Table 1. Definition of the CSF, According to Gottlieb et al. (2019)

Figure 2 highlights the user interface of the developed IT artifact asking for specific information. It consists of the CSF ("Criteria"), corresponding statements ("Statement"), radio buttons for rating and a field for entering the weight accepting only integers from 0 to 5 ("Weight"). A weighting factor is needed since the circumstances of every DV/hybrid/ETL project are unique. Therefore, besides the rating of every CSF, the importance (weight) must be ratable leading to unique recommendations for every project.

No.	Criteria	Statement	low 1	2	3	4	5	6	high 7	Weight
1	Source Data Quality	The amount of effort needed for data cleansing and validation steps is …	◉	○	○	○	○	○	○	
2	Transformation Need	The complexity of transformation steps to the target schema is …	◉	○	○	○	○	○	○	
3	Extend of Change Tracking	The extend of tracking changes in data in the source system is …	◉	○	○	○	○	○	○	
4	Source System Availability	The amount of time the data source systems are online is…	◉	○	○	○	○	○	○	
5	Computing Capacity	Source system utilization is …	◉	○	○	○	○	○	○	
6	Budget	Budget for new development is …	◉	○	○	○	○	○	○	
7	Replication Constraints	Degree of compliance and policy restrictions to replicate data is …	◉	○	○	○	○	○	○	
8	Data Model Stability	Frequency of changes in the data model of data sources is …	◉	○	○	○	○	○	○	
9	Time to Market	The maximum amount of time available until the data integration solution must be implemented is …	◉	○	○	○	○	○	○	
10	Technology Freedom	Variety of data consumers using different tools for their purpose is …	◉	○	○	○	○	○	○	
11	Agility	Frequency and degree of changes in the data strategy of the organization is …	◉	○	○	○	○	○	○	
12	Target Data Format	The complexity of the target data schema is multidimensional (high) or easily accessible with SQL (low)	◉	○	○	○	○	○	○	
13	Data Volume	The amount of data accessed in data sources is …	◉	○	○	○	○	○	○	
14	Refresh Intervals	Frequency in which data consumers need updates is … (near real-time = high; scheduled batch updates = low)	◉	○	○	○	○	○	○	
15	Application Area	The complexity of analytic purposes is … (Long-term planning, strategic = high …)	◉	○	○	○	○	○	○	

Figure 2. Score Interpretation

Once the ratings and weights are submitted, the underlying heuristic starts with checking the entries for the CSFs "Replication Constraints," "Source System Availability," and "Extent of Historization," which are exclusion criteria. If one of these factors is rated as a match for ETL or DV, the other approach is immediately excluded. In the case of replication constraints, strict policies or restrictions by law mean excluding ETL, making DV the only option, since data replication is prohibited. On the other hand, source systems with many downtimes make ETL the only option of choice, because DV requires high system accessibility. If a source system does not track changes of data by itself, ETL is also necessary here because DV does not provide a change tracking function. After this first exclusion check, based on the individually rated and weighted CSFs, a score is calculated to recommend DV, a hybrid approach, or ETL.

Score Calculation

In the first step, we inverted those CSFs (c_i) where a score (a_{ci}) of seven does not recommend DV. As described before, we included inverted cases, which have to be revoked first. Therefore, we subtract a_{ci} from 8 to achieve a comparable value (see equation 1).

$$a_{ci} = 8 - a_{ci} \tag{1}$$

In the second step, to prevent a score of four leading to a recommendation to use DV, the rating values are adjusted. Since four represents a neutral answer, recommending neither DV nor ETL, a_{ci} is subtracted by four. Thus, the scale shifted from 1 to 7 to −3 to 3.

$$a_{ci} = a_{ci} - 4 \tag{2}$$

In the third step, we normalize the weights of the 15 CSFs (w_{ci}) by dividing each weight of every CSF by the sum of all weight values of all CSFs.

$$\text{weightRatio}_{ci} = \frac{w_{c_i}}{\sum_i^{15} w_{c_i}} \tag{3}$$

Finally, the recommendation score is calculated by aggregating the results from multiplying the rating score by the normalized weight value of the 15 CSFs.

$$\text{score} = \sum_{i=1}^{15} \text{weightRatio}_{ci} * a_{ci} \tag{4}$$

Data Virtualization Score Interpretation

The calculation result is a score between −3 and +3. If a value is exactly −3 ("No DV") or +3 ("DV"), the recommended approach is not questionable. If the score is between −3 and +3, the issue that arises when trying to determine a score range recommending a hybrid solution (Hybrid) is defining the limit between ETL and Hybrid, and Hybrid and DV. Therefore, we defined two additional values, X_1 and X_2, with $-3 < X_1 < X_2 < +3$ and $X_1, X_2 \in \mathbb{R}$, as highlighted in Figure 3. These two limits cannot be fixed and must be determined individually for every data integration project or problem.

A high positive score indicates most of the rated CSFs state that it is possible, and eventually advantageous, to use DV as the only data integration method. A score is high positive when the value is in the interval $(X_2,+3]$. As a result, the conditions for applying DV are satisfied.

A high negative (low) score indicates that DV is not an option because most of the rated CSFs state that it is impossible to implement DV or there is no need to apply DV at all. A score is low when the value is in the interval $[-3, X_1)$ and ETL is recommended.

Our score only represents decision support, or a recommendation, and should not be trusted without further investigation, thus making it most useful during the early stages of a project.

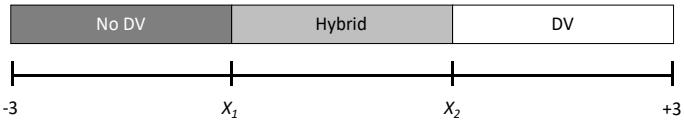

Figure 3. Score Interpretation

DISCUSSION AND LIMITATIONS

To the best of our knowledge, this is the first methodology supporting practitioners in deciding whether an ETL, a hybrid solution, or DV is the best data integration approach for their specific problem. Our methodology extends theoretical knowledge by bridging research and practice. Researchers can use the methodology to analyze or validate their concepts. Practitioners can apply the methodology in the early stages of a project to avoid overlooking essential factors influencing project success and to save time (LaValle et al., 2011; Rao et al., 2018; Schroeck et al., 2012). The rest of the discussion and the next steps are the statements, the score calculation, the score interpretation, and the complexity of our methodology.

The statements (see Figure 2) are limited to the CSFs. Even though we have conducted a systematic literature review (Gottlieb et al., 2019), we still might have missed an important CSF. Moreover, the statements are based on the findings in the literature. A content validity check is missing. In addition, there might be some more intuitive statements.

For the score calculation, we used a seven-point Likert-Scale to allow a broader differentiation than the six-point or five-point Likert-Scale. Our uneven Likert-Scale does not force the interviewee to decide between DV and no DV. However, our methodology allows analysis in detail on a spectrum. Thus, our selected scale provides additional flexibility to the individual by the ranking and weighting for their interests.

We calculated an individual score between −3 and +3. The score provides mainly an indication for a spectrum ranging from no DV to DV. Zero represents a neutral value where the decision maker is indifferent between DV and ETL. However, there is a third option illustrated by hybrid. Hybrid has all advantages from ETL and DV. The disadvantage is that both structures must be implemented and run at the same time. In addition, we use the values X_1 and X_2 to decide where the hybrid borders start and end. The borders depend on each project or problem. To provide more general values for these borders, our findings have to be extended. Therefore, we suggest applying case studies.

Furthermore, three variables must be calculated: the ranking, the weighting, and the individual borders. These three values increase our methodology's complexity.

CONCLUSION AND FUTURE WORK

In this paper, we systematically developed a fruitful IT artifact (see Figure 2) supporting decision makers, such as solution architects or project leaders, with a methodology recommending ETL, DV, or a hybrid solution with ETL and DV as an approach for data integration projects based on the identified 15 CSFs.

As a next step, we plan to validate the IT artifact with expert interviews. First, we prove the content validity of the derived statements. Second, we separate the interviews into four parts. *Part 1* measures the knowledge in the area to make sure we identify experts in the field. *Part 2* describes three cases, which must be defined in future work. *Part 3* asks experts to fill out the framework for each case. Experts should indicate the importance of each criterion and rate its expression. Thereby, reflecting on the prioritization of each criterion—the calculated score in the background is for later evaluation. Finally, *Part 4* allows experts to indicate their fitting for a solution approach with DV technology or physical data replication. From theoretical knowledge and from the literature review, we find some circumstances indicate clearly for or against a particular solution. With this assessment, we can investigate the knowledge of each expert about DV. After finishing all four parts, we can provide a validated methodology for decision support for DV.

REFERENCES

1. Bhatti, N.D. (2013) Overcoming Data Challenges with Virtualization, *Business Intelligence Journal*, Vol. 18, No. 4).
2. Bologa, A.R., and Bologa, R. (2011) A Perspective on the Benefits of Data Virtualization Technology, *Informatica Economica*, Vol. 15, No. 4), 110–118.
3. Bühner, M. (2012) *Einführung in die Test- und Fragebogenkonstruktion*, (3rd ed.). Munich, Germany, Pearson Studium.
4. Cambridge University Press. (2019) Cambridge online dictionary: time to market, in. Cambridge Dictionary Business-English from the website temoa : Open Educational Resources (OER) Portal at http://temoa.tec.mx/node/324.
5. Denodo. (2014) Data Virtualization and ETL).
6. Farooq, F. (2013) The data warehouse virtualization framework for operational business intelligence, *Expert Systems* (30, 5), 451–472.
7. Gottlieb, M., Shraideh, M., Böhm, M., and Krcmar, H. (2019 of Conference) Critical Sucess Factors for Data Virtualization: A Literature Review, *5th International Conference on Communication, Management and Information Technology (ICCMIT)*, Vienna, Universal Society for Applied Research (USAR).
8. Hevner, A.R. (2007) A Three Cycle View of Design Science Research, *Scandinavian Journal of Information Systems* (19, 2), 87-92.
9. Hevner, A.R., March, S.T., Park, J., and Ram, S. (2004) Design Science in Information Systems Research, *MIS Quarterly* (28, 1), 75-105.
10. LaValle, S., Lesser, E., Shockley, R., S. Hopkins, M.S., and Kruschwitz, N. (2011) Big Data, Analytics and the Path From Insights to Value, *MIT Sloan Management Review*, Vol. 52, No. 2).
11. Marche, S. (1993) Measuring the Stability of Data Models, *European Journal of Information Systems* (2, 1), 37-47.
12. Moxon, P. (2015) Data Integration Alternatives).
13. Rao, L., McNaughton, M., and Mansingh, G. (2018 of Conference) An Agile Integrated Methodology for Strategic Business Intelligence (AimS-BI), *Twenty-fourth Americas Conference on Information Systems*, New Orleans, LA, 1-10.
14. Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., and Tufano, P. (2012) Analytics: The Real-World Use of Big Data, *IBM Institute for Business Value—executive report, IBM Institute for Business Value*).
15. Shankar, R. (2017) Enabling Self-Service BI with a Logical Data Warehouse, *Business Intelligence Journal*, Vol. 22, No. 3).
16. Statista. (2018) Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit in den Jahren 2018 und 2025 (in Zettabyte), in. Online.
17. Van der Lans, R.F. (2012) *Data virtualization for business intelligence systems: Revolutionizing data integration for data warehouses*. Amsterdam, Elsevier/Morgan Kaufmann.
18. Vinay, S. (2012) Logical Data Warehousing for Big Data: Extracting Value from the Data!, *Gartner*).
19. Voet, M. (2018) "Data Virtualization is a Revenue Generator." Retrieved December 12, from http://www.datavirtualizationblog.com/data-virtualization-revenue-generator/
20. Yuhanna, N., and Gilpin, M. (2012) The Forrester Wave: Data Virtualization, Q1 2012, *Forrester*).