INF FAKULTÄT FÜR INFORMATIK

Databases and Software Engineering

An Analytical Model for Data Persistence in Business Data Warehouses

Veit Köppen, Thorsten Winsemann, Gunter Saake

Otto-von-Guericke University Magdeburg

May 14th, 2015

Presented by: Azeem Lodhi

Agenda

- Motivation
- Business Data Warehouse
- Decision Model
- Evaluation
- Conclusion

An Analytical Model for Data Persistence in Business Data Warehouses

Veit Köppen #1, Thorsten Winsemann *2, Gunter Saake #3

Institute for Technical and Business Information Systems, Otto-von-Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany
1 veit.koeppen@ovgu.de
3 gunter.saake@ovgu.de

* SAP
Germany
2 thorsten.winsemann@t-online.de

***Abstract*—Redundancy of data persistence in Data Warehouses is mostly justified with better performance when accessing data for analysis. However, there are other reasons to store data redundantly, which are often not recognized when designing data warehouses. Especially in Business Data Warehouses, data management via multiple persistence levels is necessary to condition the huge amount of data into an adequate format for its final usage. Redundant data allocates additional disk space and requires time-consuming processing and huge effort for complex maintenance. That means in reverse: avoiding data persistence leads to less effort. The question arises: What data for what purposes do really need to be stored? In this paper, we discuss decision support and evaluation approaches beyond cost-based comparisons. We use a compendium of purposes for data persistence. We define a model that includes objective indicators and subjective user preferences for decision making on data persistence in Business Data Warehouses.**

We develop an indicator system that enables the measurement of technical as well as business-related facts. With multi-criteria decision methodology, we present a framework to objectively compare different alternatives for data persistence. Finally, we apply our developed method to a real world Business Data Warehouse and show applicability and integration of our model in an existing system.

I. Introduction

Data Warehouses (DW), especially Business Data Warehouses (BDW), are often characterized by enormous data volumes [1]. When designing and operating a BDW, there are high requirements regarding data provision, such as performance, granularity, flexibility, and timeliness.

Besides, restrictions call for additional, redundant data. For example, materialized views and summarization levels are commonly used for enhancing speed of data access.

Performance is still the main reason for storing data redundantly, but there are several other reasons which lead to additional data persistence, as for instance design decisions, usability, or security. Such reasons are often underestimated or even not considered, but frequently motivating data storage. Yet, additional storage always requires a huge effort to guarantee consistency and limits prompt data availability. Latest technology, based on in-memory databases (IMDB), improves response times for data access and implies to enable non-redundant data access without any loss of performance [2]. Assuming that performance of data access - as the main reason for storing data redundantly - is less important in IMDB, the question arises, which data persistence is still necessary in BDW? As less persistent data means less effort, the necessity of persistence has to be questioned. In order to identify such necessities, we present a compendium of purposes for data persistence and use it as a basis for decision-making whether to store data or not. Moreover, demands on BDW are changing and evolving continuously. So, the need for existing persistent data has to be questioned occasionally. With this background, we discuss decision support approaches to define which data are needed persistently. Our approach goes beyond cost-based comparisons and also includes user preferences by using methods of multi-criteria decision analysis [3]. Note, such considerations are valid for BDW based on relational databases, too; however, the lower performance of the database motivates additional data storage anyway.

In this paper, we extend our work in [4] and present a formal methodology for evaluating persistence in BDW. With the help of a formalization of measuring and evaluation technical and business-driven indicators as well as user preferences in a holistic comparison, we decide for data persistence in an objective way. Our findings are presented with an artificial example that shows the general applicability of our approach. Furthermore, we evaluate our approach with a real-world case study, which is adapted to the systems specifics and therefore, this evaluation additionally shows applicability to for another scenario.

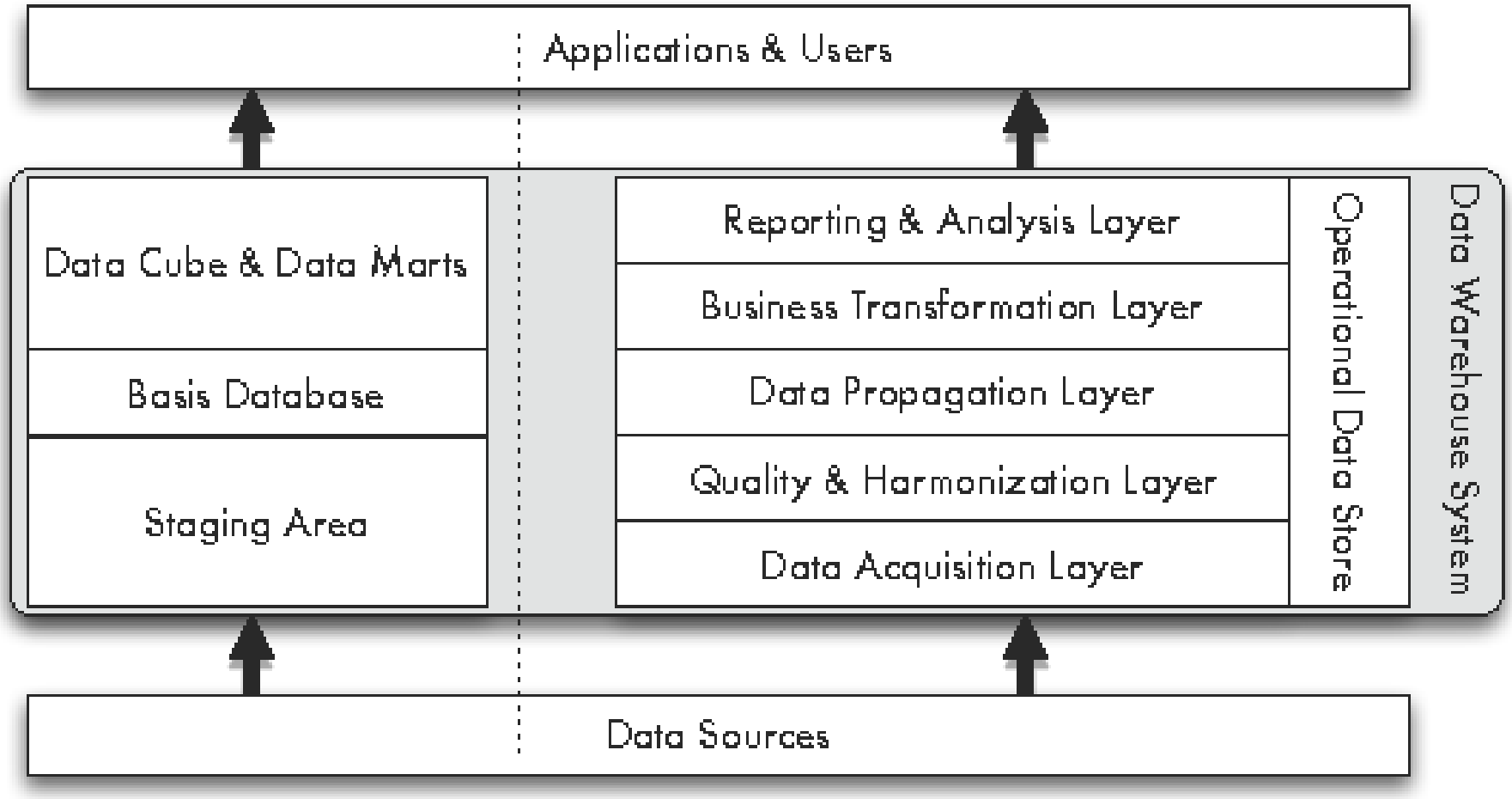
This paper is organized as follows: In Section II, we briefly introduce BDW specialties and the layered architecture. This section is based on our findings in [5] and are used as background for the domain of our decision model. In Section III, we present a compendium of reasons for data persistence in BDW and we define a process for decision support of data persistence. Both sections build the basis for our formal extension from [4] and [5] that we present in this paper. In Section IV, we define a ratio system as a basis for decision-making, enhance it with means of multi-criteria decision analysis (MCDA) methods and present an approach for an evaluation model for data persistence. In Section V, we show and discuss evaluation results of our persistence model with the help of a real-world case study. Section VI gives an overview on related work. Finally, Section VII concludes our

Motivation

- In-Memory Databases
  - Data Warehouses
  - Big Data Architectures
  - Data Lakes
- Where do we should store data?
  - Availability
  - Data Quality
  - Intermediate results
- Databases as primary storage

Business Data Warehouse

A BDW is a Data Warehouse to support decisions concerning the business on all organizational levels and it is, for instance, the basis for BI, planning, and CRM.



BDW: Layered Architecture

5 Layers

data value increase at every level

data can be used from every level

layer is not data storage

- **Data Acquisition Layer**
„inbox“ of a BDW
- **Quality & Harmonization Layer**
data integration
- **Data Propagation Layer**
Single version of Truth without business Logic
- **Business Transformation Layer**
transformation to business needs
- **Reporting & Analysis Layer**
enhancing report performance

Reasons for Data Persistency

Data acquisition: decoupling, availability, data lineage, …

Data modification: changing transformation rules, addicted transformations, complexity

Data management: constant data basis, en-bloc data supply, authorization, single version of truth, corporate data memory

Data availability: information warranty, performance

Laws and provisions: corporate governance, laws and provisions

Winsemann & Köppen (2012)

Decision Model

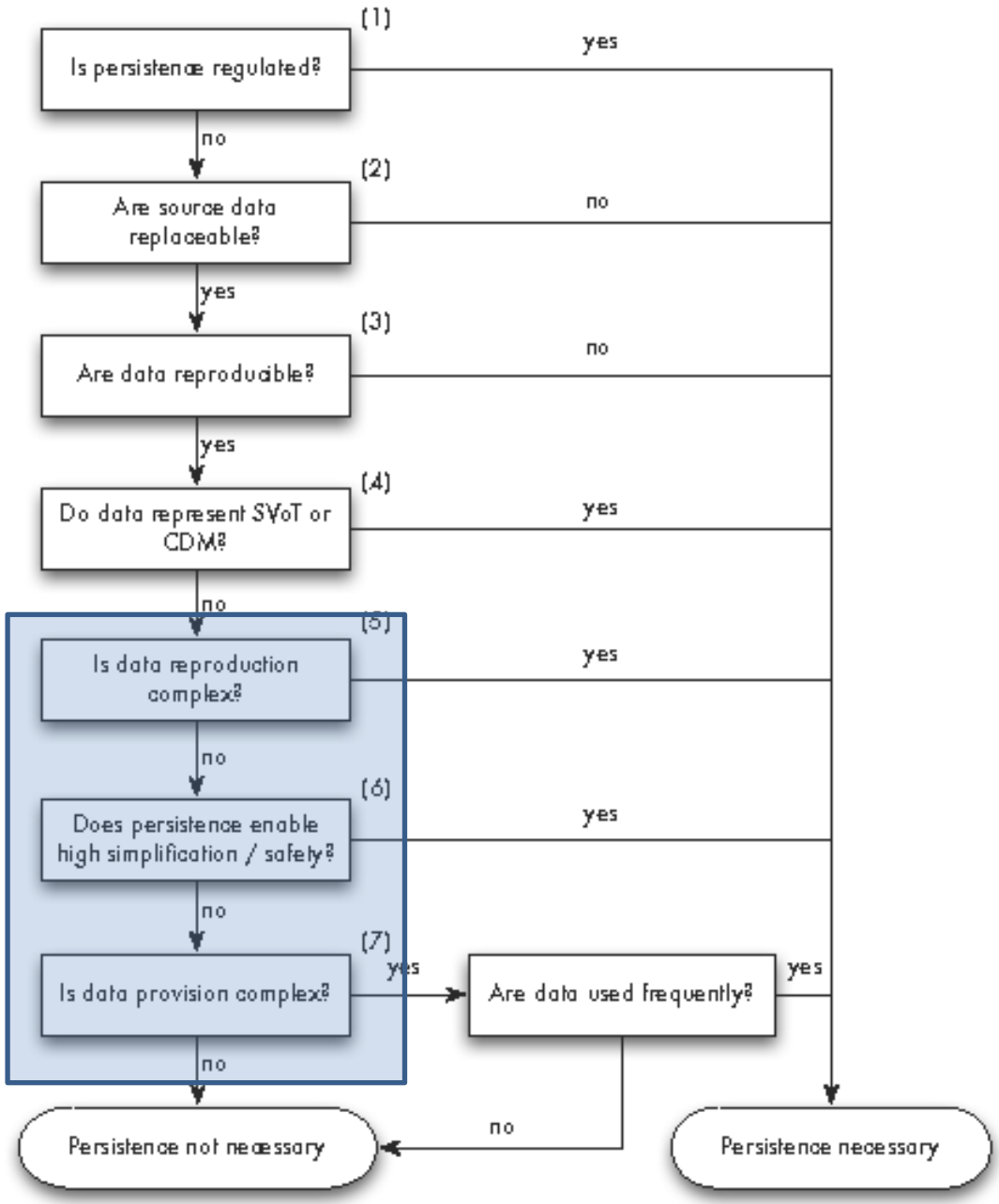
(1) - (3): mandatorily stored data

(2)- (4): „easily“ answered

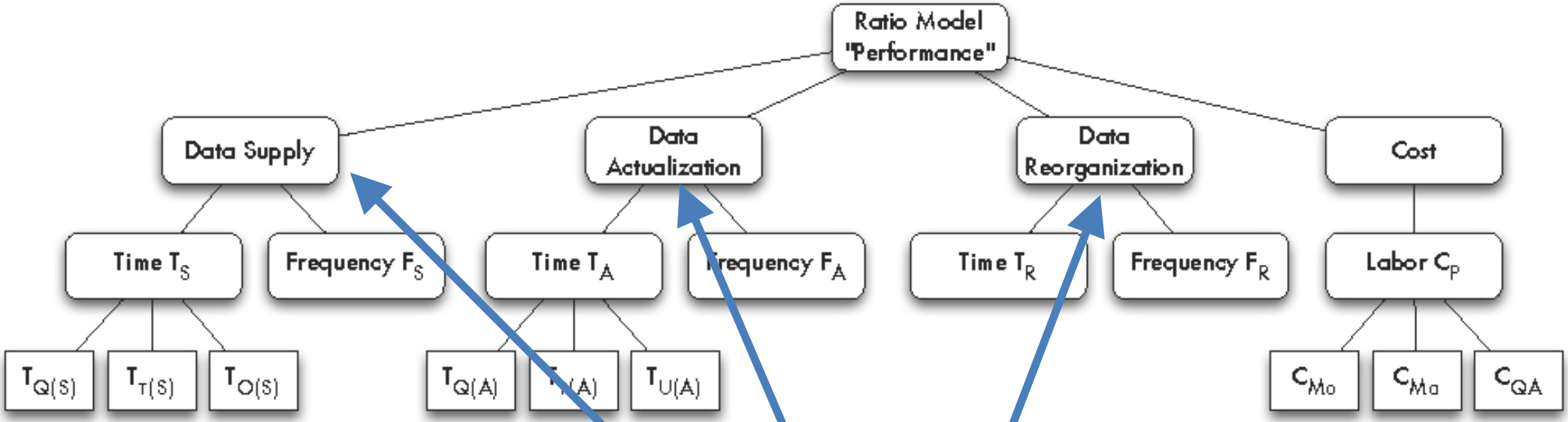
(5) - (7): fuzzy terms: complex, frequently

Focus on how to objectify

We use indicators & MCDA



Decision Model: Indicator system



Legend:
- Query (Q)
- Transformation (τ)
- OLAP (O)
- Update (U)
- Modelling (Mo)
- Maintenance (Ma)
- Quality Assurance (QA)

Measured in time units from the BDW

Decision Model: Indicator system

$$Data\ Supply = \sum_{i=1}^{Calls} F_{S_i} \cdot T_{S_i}$$

$$= \sum_{i=1}^{Calls} F_{S_i} \cdot \left(T_{Q(S_i)} + T_{\tau(S_i)} + T_{O(S_i)}\right)$$

$$Data\ Actualization = \sum_{j=1}^{Loads} F_{A_j} \cdot T_{A_j}$$

$$= \sum_{j=1}^{Loads} F_{A_j} \cdot \left(T_{Q(A_j)} + T_{\tau(A_j)} + T_{U(A_j)}\right)$$

$$Data\ Reorganization = \sum_{k=1}^{Runs} F_{R_k} \cdot T_{R_k}$$

$$Cost = C_P = C_{Mo} + C_{Ma} + C_{QA}$$

Including User preferences

Multi-Criteria Decision Analysis (MCDA)

User preferences by pairwise comparison of criteria: C_1 to C_2

1 = equal
3 = moderate
5 = strong
7 = very strong or demonstrated
9 = extreme

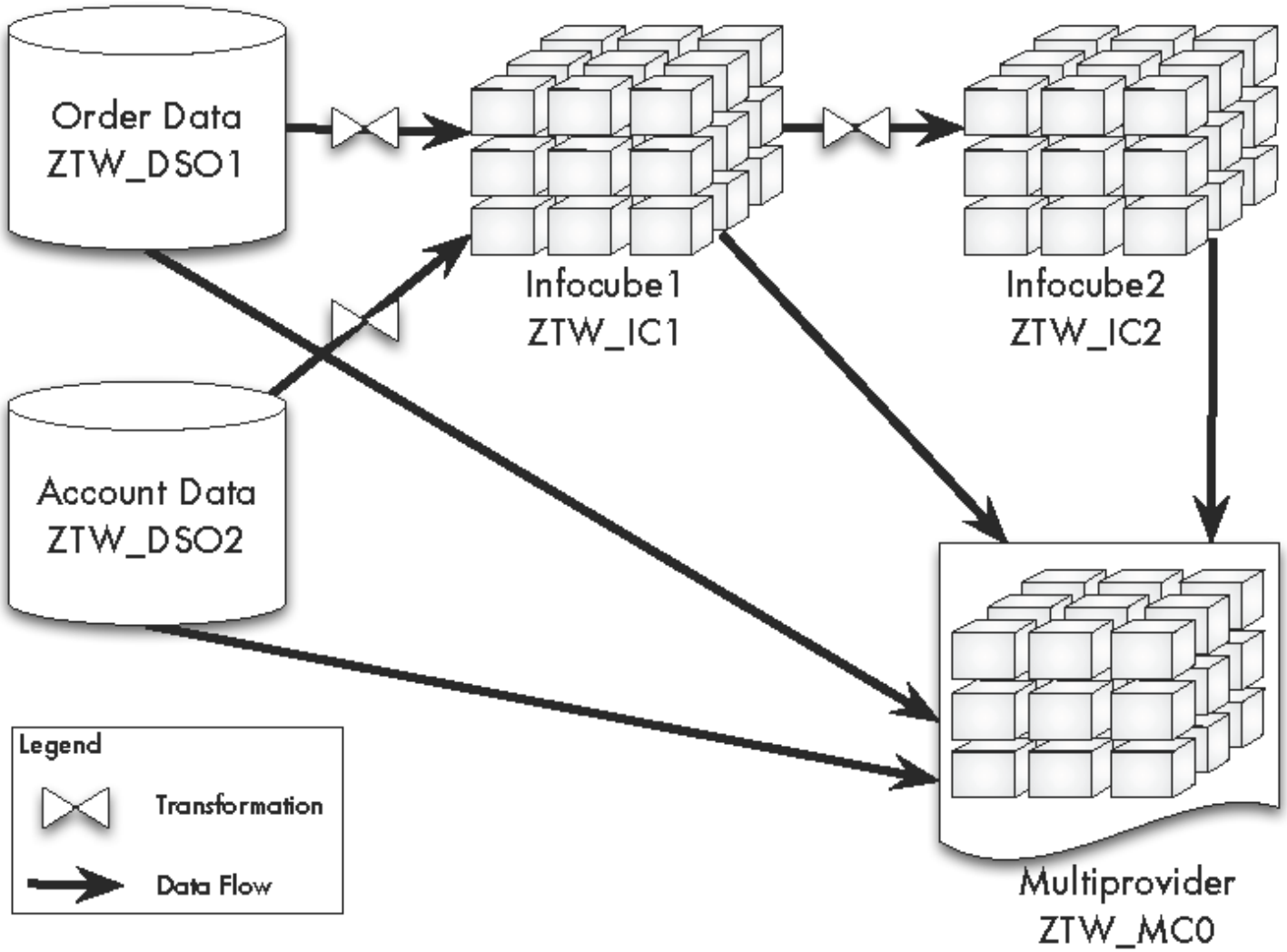
	DS	DA	DR	Cost	$\sum c$	WF_c
DS	1	6	7	5	19.000	0.44225
DA	1 / 6	1	7	1 / 6	8.333	0.19397
DR	1 / 7	1 / 7	1	1 / 7	1.429	0.03325
Cost	1 / 5	6	7	1	14.200	0.33053
$\sum\sum c$					42.962	1

Note, C_2 to C_1 is reciprocal

Result: weigthed factors (WF) per criterion

Evaluation: Scenario

SAP NetWeaver Business Warehouse, Release 7.40, with SAP HANA (HDB, Release 1.0)



Evaluation: Data elements

Order Data Store (DSO1)

Invoice Data Store (DSO2)

Infocube 1 (IC1)

Infocube 2 (IC2)

M: Measurement (detailed)
F: Fact (aggregated)

InfoObject	Type	DSO1	DSO2	IC1	IC2
Order reason	M	X		X	
Order no.	M	X		X	
Account no.	M		X	X	
Calendar day	M	X	X	X	X
Calendar year / month	M	X	X	X	X
Classification of customer	M	X	X	X	X
Customer no.	M	X	X	X	X
Region ID	M	X	X	X	X
Material type	M	X	X	X	X
Material group	M	X	X	X	X
Material no.	M	X	X	X	X
Quantity unit	M	X	X	X	X
Employee no.	M	X	X	X	X
Position no.	M	X	X	X	
Product hierarchy	M	X	X	X	X
Category	M	X	X	X	X
City	M	X	X	X	X
Sales organization	M	X	X	X	X
Distribution channel	M	X	X	X	X
Currency key	M	X	X	X	X
Number of line items	F	X	X	X	X
Order quantity	F	X		X	X
Order value	F	X		X	X
Account quantity	F		X	X	X
Account value	F		X	X	X
VAT	F		X	X	X
VAT (in %)	F		X	X	X
Open order quantity	F	X		X	X
Open order value	F	X		X	X
Returns	F	X		X	X
Value of returns	F	X		X	X
Unit price	F	X	X	X	X

Evaluation: Assumptions

Indicators are directly measured within the system

- Data Supply (6 different reports):

  20- 40 calls per day, 5 days

- Data Actualization (ETL processes):

  every 2 hours

- Data Reorganization (deletion from change logs + compression):

  Once a week

- Reports from Multiprovider → three alternatives

  DSOs, IC_1, IC_2

Evaluation: Data Supply

	Report	Average response time (s)	Frequency	Overall time (s)	Time Data Manager (s)	Time OLAP (s)	Time Frontend (s)
Data Supply DSOs	ZTW_Q01A	4.780	100	478.077	26.504	447.678	3.893
	ZTW_Q02A	2.887	100	288.668	17.088	268.557	3.024
	ZTW_Q03A	2.812	100	281.231	19.236	258.701	3.293
	ZTW_Q04A	4.020	100	402.016	19.612	378.97	3.436
	ZTW_Q05A	1.794	200	358.738	18.439	333.891	6.418
	ZTW_Q06A	2.614	200	522.850	18.157	498.225	6.464
	Overall	2.915	800	2,331.58	119.036	2,186.022	26.528
Data Supply IC1	ZTW_Q01B	4.289	100	428.905	3.222	421.953	3.716
	ZTW_Q02B	2.638	100	263.829	3.089	257.778	2.959
	ZTW_Q03B	2.491	100	249.084	3.117	242.799	3.167
	ZTW_Q04B	3.697	100	369.728	3.196	363.104	3.430
	ZTW_Q05B	1.627	200	325.352	3.560	315.897	5.896
	ZTW_Q06B	2.431	200	486.171	2.995	477.012	6.153
	Overall	2.654	800	2,123.069	19.179	2,078.543	25.321
Data Supply IC2	ZTW_Q01C	4.248	100	424.771	3.147	417.939	3.681
	ZTW_Q02C	2.636	100	263.619	2.994	257.709	2.919
	ZTW_Q03C	2.465	100	246.497	3.076	240.245	3.172
	ZTW_Q04C	3.721	100	372.125	3.189	365.511	3.428
	ZTW_Q05C	1.620	200	324.063	4.131	314.060	5.867
	ZTW_Q06C	2.437	200	487.301	2.878	478.217	6.197
	Overall	2.648	800	2,118.376	19.415	2,073.681	25.264

Actualization & Reorganization

	Process type	Average time (s)	Fre-quency	Overall time (s)
Data actualization				
DSOs	Data transfer	7.972	120	956.633
	Data activation	2.701	120	324.116
	Overall	5.336	240	1280.749
IC1	Data transfer	7.981	120	957.727
IC2	Data transfer	7.199	60	431.916
Data reorganization				
DSOs	CL deletion	22.082	1	22.082
IC1	Compression	26.717	1	26.717
IC2	Compression	24.138	1	24.138

Cost

Modelling: Two times a year

Including a new object at every element → 63 min

→DSOs: 28 min, IC_1: 49 min, IC_2: 63 min

Maintenance: 10 min per week for all

Quality Assurance: after model change: 15 min

→ DSOs: 15 min, IC_1: 30 min, IC_2: 45 min

MCDA Result

		Alternative		
Class		IC2	IC1	DSOs
Data Supply	Overall in s	2118,380	2123,070	2331,580
	Weighted utility	0,344	0,343	0,313
Data actualization	Overall in s	2670,420	2238,480	1280,760
	Weighted utility	0,234	0,279	0,487
Data reorganization	Overall in s	72,937	48,799	22,082
	Weighted utility	0,172	0,258	0,570
Cost	Overall in min per week	14,731	13,615	12,231
	Weighted utility	0,304	0,329	0,366

Evaluation: User preferences

Fast data supply & low cost

	Data supply	Data actualization	Data reorganization	Cost
Data supply	1	7	9	5
Data actualization	1 / 7	1	1	1 / 5
Data reorganization	1 / 9	1	1	1 / 7
Cost	1 / 5	5	7	1

$$w_{Data\ Supply} = 0.65, w_{Data\ Actualization} = 0.06$$

$$w_{Data\ Reorganization} = 0.05, w_{Cost} = 0.24.$$

DSOs » IC_1 » IC_2

Conclusion

- In-memory data management does not squeeze persistency out
- Decision model for data persistency
- Including user preferences with MCDA
- Can be easily adopted

Data Persistency decisions are necessary in Business Data Warehouses.

The End (?)

Thank you for your attention!

veit.koeppen@ovgu.de

thorsten.winsemann@t-online.de

References

- V. Belton and T. J. Stewart, Multiple criteria decision analysis – an integrated approach. Springer, 2002.
- J. Haupt, “The BW Layered Scalable Architecture (LSA),” SAP: Training Material, March 2009.
- H. Plattner, A. Zeier, In-Memory Data Management: Technology and Applications, 2nd ed. Springer, 2012.
- N. Roussopoulos, “Materialized views and data warehouses,” SIGMOD Rec., vol. 27, no. 1, pp. 21–26, Mar. 1998.
- O. Shmueli and A. Itai, “Maintenance of views,” in SIGMOD ’84. ACM, 1984, pp. 240–255.
- T. Winsemann, V. Köppen, “Persistence in data warehousing,” in RCIS. IEEE, 2012, pp. 445–446.
- T. Winsemann, V. Köppen, “Persistence in enterprise data warehouses” Otto-von-Guericke University Magdeburg, Technical Reports 2-2012, March 2012.