www.prudenza.nl

The next generation EDW

Letting go of the idea of a single version of the truth

Date published: august 25, 2008
Published in: Database Magazine (Netherlands)
Authors: Ronald Damhof, Lidwine van As

Enterprise datawarehouses (abbreviated as EDW) have a bad image.
They are characterised by long development processes, complex maintenance, high costs and disappointing durability.

There are two reasons for this bad image, namely a datawarehouse industry that produces architectures that do not live up to expectations, as well as market sentiment. Whereby both factors obviously reinforce one another. The datawarehouse industry should be concerned by the fact that EDW is not understood. The key is to communicate that an EDW is not synonymous with expensive, cumbersome and extended, but instead business case driven, adaptive and durable.

This article attempts to lift the negative sentiment around EDW by creating an architecture based on clear ambitions, which will enable an organisation to sustainably capitalise on its most underrated production factor: data. An architecture that differs fundamentally from those realised by datawarehouses during the past decade. This article does not primarily concern the business justification of an EDW, but instead highlights the ambitions of an EDW and the resulting choices in terms of architecture. This article has, both in terms of ambitions and architecture, been given the pretentious title; ' next generation'.

Ambitions
The first priority in setting up the architecture and infrastructure of an EDW is to clearly define the ambitions. It should not get bogged down in a quagmire of technology. Instead, the ambitions should be worded in such a way that they are understood by the organisation, and backed by its management. Obviously, these ambitions must be supported by a detailed business case model, clearly setting out the potential returns for the organisation and the payback time.

The ambitions for the architecture and infrastructure of the new generation EDW are:
Complete traceability of data. Data stored in the EDW and intended for use by the organisation must be traceable to the original source. Additionally, the data must comply with modern compliance requirements.

- Meaningful data. All data in the EDW must be accompanied by meaning in the sense of definition, owner, business rule, domain values, etc.

- A degree of unbundling between data and the operational system landscape. Not every change should impact heavily on the EDW with consequences for the end user. If properly modulated, data structures in the EDW can be significantly more stable, compared with constantly changing organisational processes and their impact on the operational system landscape.

- Data quality and business rules are the responsibility of the business, not the domain of the EDW team or the ICT department. EDW must, however, provide for a clear separation of these responsibilities.

- Scalability all its forms. EDW is, by definition, an incrementally growing architecture. More data, more users, more applications. The architecture must enable the infrastructure to grow in line with requirements. Scalability concerns both scalability in resourcing and in technical terms. An example is the aspiration to have maximum management say in the performance of the data logistic processes. An EDW architect must be able to assert within his organisation that increased resources will lead to faster data loading. For if that is not the case, processes will slow down. Zero-update strategies in combination with parallelisation are essential in this respect.

- The EDW must enable the business to confront and integrate data. The new generation EDW requires this to be done ‘just-in-time’, not in advance therefore. 'Just-in-time’ integration means that we only integrate when the business requires it. In this, two principles are essential; business rules are implemented downstream (in the direction of the end user), and a distinction - up to physical level - is made between ‘Facts and Truth’. There is no ‘Single version of the truth’.

- The EDW must be able to achieve ongoing reductions in the throughput times of requested information products.

- Data from datawarehouses is increasingly acquiring an operational character. Data from the operational system landscape must be delivered ever faster to the end-users (downward trend in data latency). The new generation EDW must be prepared for a high-load mixed workload.

- The datawarehouse must also comply with contemporary requirements made of software engineering. An important and very useful indicator is provided by the five levels of the Capability Maturity Model (CMM) as drawn up by the Carnegie Mellon Institute. A datawarehouse should have a minimum ambition of level 4.

- Robust use of the operational system landscape is often a prominent ambition in larger data-intensive organisations. No doubt, many will ask: how can the EDW contribute to this? Systems have - often in an uncoordinated fashion - acquired a supplementary role. Not only must they support the primary process, all too often

they are also burdened with many ad hoc questions and numerous regular datasets for various (external) users. Never mind the problems surrounding meaning, performance, scalability and management, this is an undesirable situation. The new generation EDW enables the organisation to effectively (properly and fully) access sources and distribute them in a manner that satisfies all the aforementioned ambitions.

Besides the above, an EDW must of course provide optimal performance and user-friendliness, while BI services (reporting, analysis, mining, etc.) must be formulated to facilitate the valuable deployment of data. Below, I explain how these ambitions can be realised by taking a different approach to data and data logistics. The principles of Dan Linstedt with his Data Vault architecture play a very prominent role in this.

The data: fact and truth

Up to now, datawarehouses had the objective, implicitly or explicitly, to create a single version of the truth for their users. Companies struggling with inconsistent information drawn from separated datasilos welcomed the datawarehouse as the ideal solution for the underlying disarray. Gradually, however, they discovered that this was little more than an attempt to create an IT solution for what was, in essence, a business problem; for if the organisation does not have a company-wide uniform vision of its data, the datawarehouse is unable to impose it. After all, there will always be users who (voluntarily or out of necessity) adopt a deviating view of the available data.

By enforcing **'one version of the truth'** in the datawarehouse, those users are left out in the cold. Other considerations also started to play a role. Since the start of the century, companies have been confronted with the need for compliance with regulations such as Sarbanes-Oxley and Basel-II. This requirement is at odds with the storage of interpreted, processed data. If the original data is no longer traceable (and the datawarehouse is usually the only place where the history is stored), we lose sight of the data 'trail'; data is no longer traceable and the datawarehouse is not compliant. The new generation datawarehouses lets go of the idea of a single version of the truth, adopting instead the 'truth is in the eye of the beholder' perspective. The goal is therefore to create a system, i.e., a '**single version of the facts**', which is open to individual interpretation. The single version of the facts offers scope for multiple versions of the truth. Whatever interpretation users apply, they can always be attributed to the pure facts - meaning that the datawarehouse meets the compliance requirement.

Data quality

The wish to maintain a system of fact also leads to a different view of the quality of data. Point of departure is that all data should always be loaded, irrespective of its low quality: 100% of the data 100% of the time. Justification is found in the argument that 'low quality' is also a matter of interpretation; what may be unusable for some users, is more than adequate for others. What's more, this concerns data stored on an operational system that may form the basis for decision making. It is therefore up to the individual user groups to determine whether the data is good enough to use - and to specify how this quality should be improved in their version of the truth.

Integration and cleaning only take place after the datawarehouse

The T in ETL stands for Transformation and the question is where the bulk of the transformation should be positioned in the EDW. In the traditional datawarehouses, it is positioned between the staging and the EDW (see Figure 1). An often-heard argument is that the datawarehouse should contain clean data; the concept of 'a version of the truth' often relates to this part of the datawarehouse.

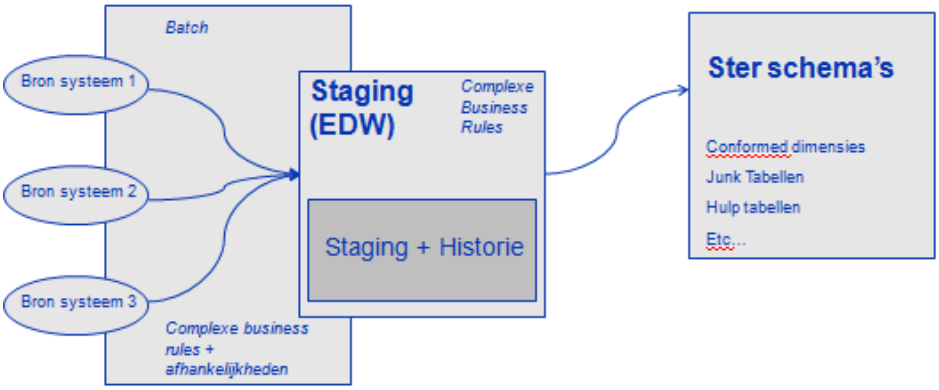


Figure 1: Traditional architecture, business rules before the EDW.
(Source: Dan Linstedt - The business of Data Vault modelling)

In time, this will give rise to problems:

- increasing difficulty/impossibility to trace the data;
- increasing difficulty to scale the ETL or retrieve batch windows;
- projects are increasingly taking longer to complete, becoming more complex and requiring ever more expensive resources;
- diminishing flexibility of the datawarehouse architecture, i.e., its ability to adjust to ever-changing requirements of the business;
- implicit history use assumptions appear to be incorrect;
- testing, especially regression tests, become increasingly difficult.

The result is a datawarehouse that becomes outdated relatively quickly, thereby losing its ability to support the business with high-quality and timely data.
It is wrong to position Transformation ahead of the EDW. In new generation EDW, Transformation should always be positioned after the EDW (see Figure 2). Preferably as near as possible to the end-user and his/her requirements. Ideally, the Transformation should take place at the instance that the user requests it. That is often not possible, however, primarily due to (technical) performance reasons.

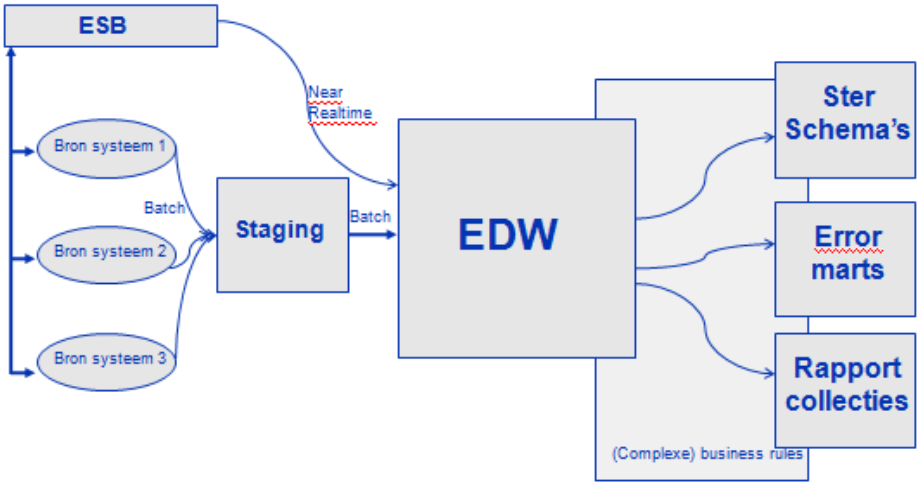


Figure 2: Fundamental architecture; Business rules after the EDW
(Source: Dan Linstedt - The business of Data Vault modelling)

An important architectural concept of the next generation EDW is that data is not integrated when the business does not require it. When the business so requests, the data is integrated between the EDW and the datamarts. The EDW is itself filled with one-on-one data from the source systems that can be loaded in a parallel, asynchronous and largely standardised manner.

Parallel loading EDW: the EDW must be modelled according to the strategy of 'zero-updates'.

Asynchronous loading: everything delivered in the staging (work units) must be accepted for direct processing.

Standardised: the EDW is modelled according to a standard method (Data Vault) with a limited number of entity types (Data Vault has 3). Every entity type exhibits uniform loading behaviour. This level of standardisation opens the door to ETL generation. It is important that ETL suppliers recognise this development and support it with either standard ETL templates/mapping or a form of model-driven datawarehousing.

As a result, loading data from the source to the EDW becomes a ‘conveyor belt process’ with automated efficiency and effectiveness. Practically any request made by the business can be met by entering the produced data intermediates into this automated process.

Data Vault as enabler

The architecture of the Data Vault fits in seamlessly with the sketched ambitions and architecture. To the extent that datawarehouse/BI professionals are familiar with Data Vault, it is generally as a modelling method (readers wanting further clarification are referred to the excellent article by Maarten Ketelaars on this topic in DB/M7, 2005). It is a pity that Data Vault is often incorrectly understood, explained and applied. The uniqueness of this method lies not so much in its chosen modelling, but rather in the manner in which it supports the architecture principles, as described in this article. In fact, Data Vault is ideal for facilitating such points of departure as ‘100% of the data 100% of the time’ and ‘asynchronous loading’. I will explain this using a few examples.

Example: problems with RI.

All the sales transactions are historically recorded in the EDW of an imaginary company, see Figure 3.

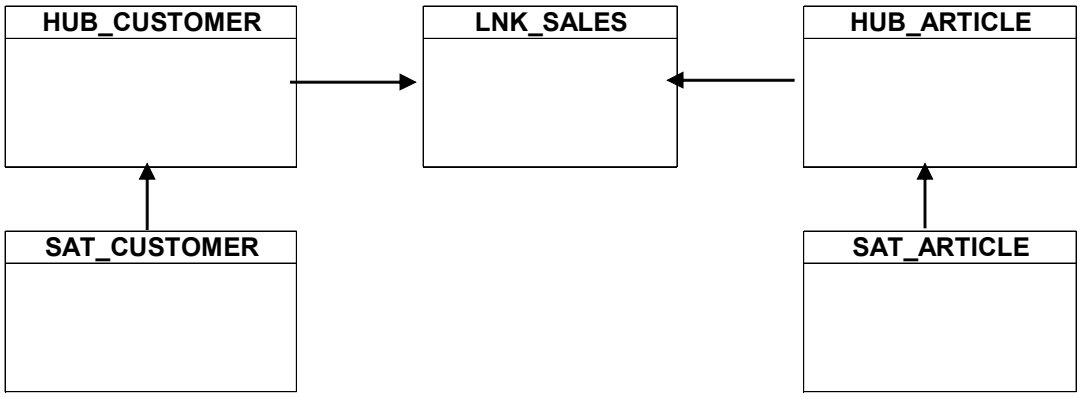


Figure 3

The CRM system of the company is the master source for client data; but in order not to hinder the sales process, sales staff can issue own client numbers, so that they can register new clients in the sales system. These client numbers are then later registered in the CRM system.
Sales transactions supplied to the EDW can therefore include an unknown client reference.

On 15-07-2008, the loading process receives the file IMP_SALES, and ascertains that client number DNZ999 is not yet present in HUB_ CUSTOMER, see Figure 4 and 5.

HUB_CUSTOMER			
SURR_ID	LOAD_DTS	CUSTOMERNR	RECSRC
1	01-01-2007	ABC123	CRM
2	01-01-2007	XYZ456	CRM
3	25-06-2008	MNO789	CRM

Figure 4

IMP_SALES					
ARTIKELNR	CUSTOMERNR	COUNT		PRICE	TRXDTS
...	...	...	...		...
1904X5	DNZ999		5	5,95	15-07-08
...	...	...	...		...
...	...	...	...		...

Figure 5

In a traditional environment, we would drop this transaction record; in the Data Vault, on the other hand, we first make a new HUB_CUSTOMER record, see Figure 6.

HUB_CUSTOMER			
SURR_ID	LOAD_DTS	CUSTOMERNR	RECSRC
1	01-01-2007	ABC123	CRM
2	01-01-2007	XYZ456	CRM
3	25-06-2008	MNO789	CRM
4	15-07-2008	DNZ999	**SALES**

Figure 6

Subsequently, the transactions can be loaded problem-free in LNK_SALES – we can therefore load all supplied data.

As the client is not yet known in the master CRM system, there is no descriptive data available, and no satellite can yet be generated for the new client, see Figure 7.

SAT_CUSTOMER			
SURR_ID	LOAD_DTS	CUSTOMERNAME	RECSRC
1	01-01-2007	ABC VOF	CRM
1	01-06-2008	ABC BV	CRM
2	01-01-2007	Xyz corp	CRM
3	25-06-2008	Acme Inc	CRM

Figure 7

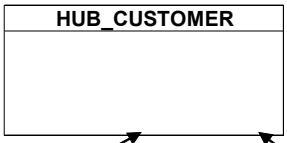
After a few days, the CRM system is updated, client DNZ999 is loaded and supplemented with name, address and other descriptive data. No exception routines or updates are now required in the EDW, the client data can be added through the regular process in SAT_CUSTOMER, see Figure 8.

SAT_CUSTOMER			
SURR_ID	LOAD_DTS	CUSTOMERNAME	RECSRC
1	01-01-2007	ABC VOF	CRM
1	01-06-2008	ABC BV	CRM
2	01-01-2007	Xyz corp	CRM
3	25-06-2008	Acme Inc	CRM
4	17-07-2008	DeNieuweZaak	**CRM**

Figure 8

Example: technically poor data quality.

The Data Vault also offers an integrated solution for data that is of such poor quality that it cannot be used in the datawarehouse environment. Examples include excessively long values, numbers not consisting of figures, deviating date formats, etcetera. In the Data Vault, these values are (with the user's permission) converted into default values:
ZERO for numbers, a default value (e.g. 1-1-1780) for dates. However, the original value is not lost during the conversion, it is stored in a special error satellite, see Figure 9.



SAT_CUSTOMER		
SURR_ID	CUSTOMERNAME	CUSTOMER_SINCE
1	ABC VOF	30-12-2007
1	ABC BV	30-12-2007
2	Xyz corp	30-12-2007
3	Acme Inc	**01-01-1780**

SAT_CUSTOMER_ERR		
SURR_ID	CUSTOMERNAME	CUSTOMER_SINCE
3		**25-13-2008**

Figure 9

The error satellite contains the same fields as its regular ‘brother’, but instead has a character datatype; every possible incoming value can thus be stored there.
The problem records from the error satellite can be passed on to an Error Mart, so that users can view and assess them, and make changes in the source system if necessary.
Quality assurance in the Data Vault has thus become part of the architecture and the regular datawarehouse process.

Conclusion

The next generation EDW has two fundamental architectural characteristics: a distinction is made between facts and truths at conceptual, logical and technical level; the level of data integration is determined by the end-user and, if technically possible, is carried out at the time of request.
What does that mean for the architecture and the datawarehouse process? A radical change. The central datawarehouse is no longer a clean, integrated data collection; it serves solely as a storage space for facts. Integration and cleaning only take place after the datawarehouse, when the datamarts are made. Tailored to meet user needs, it reflects the user's view of the facts and thus the user's truth.

While organisations have raised their requirements of datamanagement (especially in terms of compliance and scalability) during the past 10 years, the datawarehouse industry has failed to keep pace. Using the architectural principles outlined above, the industry may be able to realign itself with the ever higher demands that organisations make of datawarehouses.
Users seeking methodological support will appreciate the Data Vault as an excellent framework for the next generation EDW, one that will be better at meeting user requirements and reversing the negative sentiment around datawarehousing.

Ronald Damhof (Ronald.damhof@prudenza.nl) is Information Management Architect and Certified Data Vault Grand Master

Lidwine van As is consultant at Grey Matter and is certified Data Vault modeller