The Evolution of the Data Warehouse Systems in Recent Years

Introduction

Although data warehouses are used in enterprises for a long time, they have evaluated recently. In the last twenty years the term data warehouse was used thousands of times in academic articles. The key issues of the data warehouses were mostly described in two books: “Building the Data Warehouse” written by W.H. Inmon [Inmon, 2005] and “The Data Warehouse Lifecycle Toolkit” that was written by R. Kimball [Kimball et al., 1998]. The formal definition of the data warehouse mostly used in academic papers is: the data warehouse is a repository that has four attributes: subject-oriented, nonvolatile, integrated and time-variant. That definition comes from W.H. Inmon while the rules of slowly changing dimensions are mostly referred to R. Kimball.

The article presents the conditions that had significant influence on the evolution of the data warehouse environment. It does not divide data warehouse term into various form of this system such as: enterprise data warehouses, data marts, operational data warehouse etc. It digests the definitions that have been used to describe data warehouse since 1991, when the book “Building the Data Warehouse” was published.

The article is divided into six general parts – in the first part a theoretical background was presented to classify the most common used definitions of the data warehouse. In the second and third part the key issues of the data warehouse evolution were presented, such as treating unstructured data. The fourth part of the paper presents the analysis of the data warehouse attributes, whether they are still valid and useful. In the fifth part the paper solutions to mitigate data warehousing weaknesses were presented. Last part shows conclusions.

The thesis of this article is as following: the role of the data warehouse systems has changed recently. One of the business environmental changes that has a big impact on data warehouses is making analysis based on unstructured data. In 1991 when the data
warehouse term has been defined by W. H. Inmon, unstructured data were not concerned in the data warehouse. In Internet era, which grows immediately in the turn of 20\textsuperscript{th} and 21\textsuperscript{st} century, there is a need of including these data. This is one of the changes in the data warehouse, but not the only, as it was written in further parts of this article.

1. Theoretical background

As mentioned in the introduction there are several definitions of the term data warehouse, but the most often used in academic papers are based on W.H. Inmon and R. Kimball.

There are lots of books that use these definitions. For instance definition based on Inmon was used in [Humphries, 1999, p. 34], [Abramowicz et al., 2000, p. 7-8], [Ponniah, 2001, p. 23-24], [Malinowski, Zimanyi, 2009, p. 3], [Todman, 2001, p. 32-33]. This well-known data warehouse definition (subject-oriented, non-volatile, integrated and time-variant) concerns the characteristics of the data in the warehouse not the data warehouse system.

That is why there are books that combine the definition of the data warehouse based on Inmon with another definition, for instance based on Mattison [Abramowicz et al., 2000, p. 7-8]. It is possible because Mattison defined data warehouse as the objectives of the system: “…whose primary purpose is to extract information out of legacy systems, and make it usable to business people, in the support of their efforts to reduce cost and improve revenue”.

Based on Hammergren “a data warehouse is a repository of subjectively selected and adapted operational data, which can successfully answer any ad hoc, complex, statistical and analytical queries. It is situated at the centre of a decision support system (DSS) of an organization and contains integrated historical data; both summarized and detailed information – common to the entire organization…” [Prabhu, 2006, p. 1].

There are also several books that use a bit modified definition based on R. Kimball, for instance: a data warehouse is a large analytical database which derives its data from a variety of production systems and is structured for querying, reporting and analysis [Khan, 2003, p. 8].

The definition based on Oracle documentation is as follows: a data warehouse is a database containing data from multiple operational systems that has been consolidated, integrated, aggregated and
structured ... to support the analysis and decision-making process of a business [Hobbs, 2005, p. 4].

Several researchers defined the data warehouse based on its functional aspects. In that sense the data warehouse is an information system that: provides an integrated and total view of the enterprise, make the enterprise’s current and historical information easily available for strategic decision making, makes decision support transaction possible without hindering operational systems, renders the organization’s information consistent, presents a flexible and interactive source of strategic information [Ponniah P., 2001, p. 13].

But the data warehousing systems were used before 1991 when the formal definition of the data warehouse was established. The first step towards data warehousing was the fixed presentation of data. The aim was to build a system on top of automation and controlling applications, an example of the first warehouse is based on Wal-Mart [Westerman, 2001, p. 4] or IBM Ireland [Januszewski, 2008, p. 21].

To conclude, data warehouse is still defined in three ways:
- concerning data, usually based on Inmon,
- as a system,
- based on its functionality.
The most flexible definition is based on W. H. Inmon but it is very often supplemented with functional or system definition. One of the definitions concerning data warehouse as a system is based on R. Kimball.

2. Data warehousing concepts in recent academic papers

An important step is to analyse the current practical aspects of the data warehouse based on recent academic articles, reviewed by universities authorities. Table 1 shows the number of articles containing the term “data warehouse” in its keywords or titles in well-known research papers databases in last five years.

<table>
<thead>
<tr>
<th>Table 1. Number of articles concerning data warehousing issues in selected databases in January 2008–April 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name of the database</strong></td>
</tr>
<tr>
<td>EBSCOhost (Business Source Complete)</td>
</tr>
<tr>
<td>Web of Science</td>
</tr>
<tr>
<td>SciVerse</td>
</tr>
</tbody>
</table>

Source: Own elaboration.
The next step was to analyse the context of these articles. It can be found that these articles mostly concern data warehouse designing issues, such as integrating issues, data cleansing, conceptual designing (including ontology-based design), security aspects, reporting issues, metadata managing, business intelligence and data mining mechanism. There are also several articles based on case studies that describe challenges of the data warehouse implementations in different enterprises as well as surveys results of the data warehouses implementations in business organizations.

There has also been developed a concept of the streaming data warehouse. The key issue is to combine the traditional data warehouses and data stream systems [Golab et al., 2012, p. 1092-1093]. This concept was mentioned several times in articles, concerning different aspects of its functionality, including OLAP queries and scheduling.

Another concept of the data warehouse that was tackled in research papers is real-time and near-real-time data warehouses. A real-time data warehouse incorporates operational data changes in real time [Chen et al., 2011, p. 1150]. It collects a large number of streaming data. These warehouses are mostly used in applications such as IP network monitoring, online financial trading or credit card fraud detection [Bateni et al., 2011, p. 757]. The difference between near-real-time and real-time warehouse is that near real-time mechanism can identify whether the data warehouse needs to be updated in real-time or near real-time, which is called right time. The key issue is not to update the warehouse directly after the evidence occurred, but to verify the quality of the data based on several measures, such as frequency of the data feed.

There is also a concept called Active Data Warehouse. It is a warehouse that is refreshed online and achieves a higher consistency between the stored information and the latest data updates [Polyzotis et al., 2008, p. 976-977].

The conclusion is that the data warehouse is still a strong research area for scholars and scientists. New concepts of the data warehouse are developed and there is still a necessity to improve them.

3. Key issues of the data warehouse evolution

What has been written in previous parts of this article, analyzing academic papers in leading journals on IT and management, leads to the
The Evolution of the Data Warehouse Systems in Recent Years

conclusion that one of the most common used definition is based on W.H. Inmon’s book titled “Building the Data Warehouse”.

There were lots of academic and brief news articles concerning data warehouse evolution. In these papers it can be read about functional aspects of this system, such as that enterprise data warehouse is the next logical step in the data warehouse evolution chain [Sinn, 2002, p. 31-32]. It was also written that the data warehouse evolution is about changes and means to handle occurring changes [Papastefanatos et al., 2008, p. 444].

But one of the key issues of the data warehouse evolution is related to adapting the warehouse to changing business environment. Using agile methodologies such as Scrum and eXtreme Programming (XP) seems to be the future of the data warehouse [Hughes, 2008, p. 1-3]. As we need data quickly and there is a need to increase the functionality of the data warehouse, the only way is to use a methodology that will allow us to build such systems very fast and in efficient way. The challenge is to adapt data warehouses implementation methodology to be more flexible which can be done by using agile methodologies.

In this analysis it is also not possible to omit the key position in the data warehouse designer’s library which is “Building the Data Warehouse” written by W.H. Inmon. This book had several editions and the fourth edition, that was printed in 2005, contains the following new topics, important in today’s data warehouse environment:

- Operational Data Store – the place for high-performance response time processing;
- unstructured data – the idea was to link structured and unstructured data by text using a probabilistic match, a themed match;
- near-line storage that allows to extend data warehouse to infinity, because it allows storing the data in the low-performance and therefore inexpensive disk storage [Inmon, 2005, p. 144, 308-310, 429].

In 2006 a new approach called Data Warehouse 2.0 has been developed by W. H. Inmon and others [Inmon et al., 2008, p. 9-11]. There are several differences between traditional data warehouse and the concept named DW 2.0. But the most important difference that it relies not only on the structured data but also on unstructured, such as text documents. The Data Warehouse 2.0 should be treated as the new project, comparing to the previous written in the 4th edition of the “Building the
Data Warehouse” that has just some proposals of matching and linking unstructured and structured data.

To conclude this and previous part of the article, the following issues had the biggest impact on data warehouse evolution:

- need to store unstructured data, what is the result of increased number of data sources in enterprises, also adapting to Big Data mechanism should be considered,
- using agile methodologies as end users wanted faster implementation of the system to fulfil their changing requirements as well as users become more sophisticated in their requirements,
- streaming and real-time data warehousing that need to create a high reliability and availability environment.

4. Attributes of the modern data warehouse

4.1. General view

As it was mentioned in the introduction, based on W. H. Inmon the data warehouse is a system with four attributes:

- subject-oriented,
- non-volatile,
- integrated and
- time-variant.

In this chapter we will try to focus on changes within these attributes in recent years.

4.2. Subject-oriented

Based on W. H. Inmon subject-oriented means that the data is concentrated on subjects, such as customer, policy, premium, claim in insurance company, not on applications as it is in operational systems [Inmon, 2005, p. 29].

It depends on the way subject-oriented are considered but modern warehouses still concentrates on various subjects. It is even when the approach based on variable is used. The variable is a dimension table with attributes such as the name of the variable, subject, description etc. Each variable may concerns different subject but it is still subject-oriented, not operational storage.

But to conclude we can say that the data warehouse is still subject-oriented and there were no changes in this matter.
4.3. Non-volatile data

In Inmon we can read, that non-volatile means that once the data is loaded into the warehouse it cannot be changed [Inmon, 2005, p. 31-32]. Based on the research made in Polish organisations that use data warehouses, we can read that data are sometimes deleted, especially when there is a new source with better quality or there is a decision that the data will not be used anymore in the warehouse [Maślankowski, 2008, p. 45-46]. Another issue is that data in the data warehouse sometimes have to change, when there was a mistake or newer data is available. In particular when the data warehouse is based on macroeconomics data, such as inflation rate or GDP estimates that are usually corrected in a long period of time. The solution for this problem is to use versioning mechanism – to store different versions of data (newer or older), but for some reasons it is too complicated and it is easier to update the data. There are also several new approaches in handling the issue of updating the value of a dimension or a fact [Papastefanatos et al., 2008, p. 444].

Based on these facts, it can be said that there was a need to precise the meaning of non-volatile attribute, but it is not the key issue of the data warehouse evolution.

4.4. Integrated data

Integrated data means that the data is fed from multiple, disparate sources into data warehouse. Then the data is converted, reformatted and aggregated as well. As the result the data warehouse has a single corporate image [Inmon, 2005, p. 30].

It is a general knowledge that the data gathered in the data warehouse must be integrated. It is not possible to load data into one repository of the warehouse that are not integrated, considering the warehouse schema with common dimensions. However there were some concepts of using unstructured data in data warehouses [Inmon, 2005, p. 305-330]. Although R. Kimball clearly wrote in his second edition of the book “The Data Warehouse Lifecycle Toolkit” that the data in the warehouse repository should be structured [Kimball, Ross, 2002, p. 7-8], the necessity of including unstructured data as a source was considered by W. H. Inmon and in the 3rd edition of the book “Building the Data Warehouse”, that was published in 2005, Inmon showed how to store unstructured data in the data warehouse. In the same time W. H. Inmon made the new concept of the data warehouse called DW 2.0. In this concept the repository of the data warehouse was divided into two parts
to store data: structured and unstructured. Significant role of unstructured data is the result of the growing of social networks on the Internet.

Therefore the main evolution in the data warehouse integrated data issue is to integrate both structured and unstructured data.

4.5. Time variant

Time variant is an attribute of the data warehouse that was defined by W. H. Inmon as “every unit of data in the data warehouse is accurate as of some moment of time” [Inmon, 2005, p. 32].

Each data warehouse designed in multidimensional model contains data that referred to a specific period of time or at least specific time stamp. In that matter nothing has changed since the data warehouse has been defined. There is no possibility to deploy a data warehouse environment in modern IT data warehousing systems such as Microsoft SQL Server or Oracle without having time dimension in warehouse schema.

5. Current trends to mitigate traditional data warehouse’s weaknesses

Traditional data warehouse has several weaknesses that can be mitigated by new improvements in its design method. Issues to improve in data warehousing implementation can be divided into following groups:
- managing the changes in the data warehouse to reflect changing business environment,
- dealing only with structured data, whereas more important are also unstructured data,
- implementation agility.

The first issue is to design the data warehouse to be flexible enough to make changes in its subject and object scope without necessity of modifying its schema. For instance adding new data source to the repository can be one of these changes. Although lots has been done in this area, e.g., Ralph Kimball has developed a concept called slowly changing dimensions and William H. Inmon published the enhanced version of the data warehouse called DW 2.0, there are still some lacks in this matter. One of the ways to handle this issue is to use a variable approach in designing a data warehouse [Maślankowski, 2013]. The approach based on variables focuses on each business fact as a variable that can be simply described. The goal of the variable approach is to add
a new dimension table named Variable. This table should have at least few attributes (dimensions) such as the name of the variable, its description, group and subgroup, the possible ways of calculating (aggregations, possible calculations), methods of retrieving the result, data source (if possible) and others if necessary. The aim of the variable dimension is to add new phenomenon from business environment, not only new product but also some calculations, mathematical or statistical computations and similar.

The second issue is that the traditional data warehouse gathers only structured data. However, what has been mentioned before, W. H. Inmon published the concept of the new data warehouse called DW 2.0 that is in fact the repository to integrate unstructured and structured data together [Inmon et al., 2008]. This approach has been positively evaluated and was one of the subjects tackled by several researchers. There were also several different approaches to resolve this problem, e.g., a document warehouse which was a concept of gathering unstructured documents and as a prototype was developed in late 90s. Its enhanced metadata, to better maintain and manage all unstructured documents, based on keywords and content based retrieval mechanisms as well as cross-document relations [Tseng, 2005, p. 55-56], can be used not only to gather and analyze text documents, but also to store and manage multimedia files [Ishikawa et al., 1998, p. 90-91].

The third issue concerns the agility of its implementation. Based on the research made by IDC the time needed to implement the data warehouse is usually more than 18 months – this concerns 65-70% of DW implementations [Bach et al., 2012, p. 147]. It was confirmed by the study on data warehousing implementations in Polish organizations [Maślankowski, 2008, p. 44-45]. Therefore there is a need of minimizing the total time of the data warehouse implementation as well as improve its flexibility, especially by using agile methodologies. There are some elaborations of the possibilities in implementing data warehouse by using different agile methodologies, such as XP, RUP, Scrum or Lean Agile [Sandler, 2010, p. 20], but they are still not widely used.

Although this article focused only on these three issues of modern data warehouses, in literature we can find several different aspects of lacks of the data warehouse concerning its data. Firstly, the hierarchies are represented only based on balanced trees, what differentiated data warehouse environment from the real world where hierarchies are
unbalanced. Secondly, data warehouse has not been still adapted to the spatial data. Thirdly, inclusion of the temporal data must be concerned not only in the fact table, but also by dimensions and hierarchies [Malinowski, Zimanyi, 2009, p. 1-2].

Nevertheless, data warehouse proposed by W.H. Inmon and R. Kimball is still the most appropriate system for analytical purposes. The key issue is to mitigate the weaknesses by supporting the warehouse with the new approaches that were presented above.

**Conclusion**

Although lots has been written about data warehouses there is still a need for improving existing methodologies of its implementation as well as enhancing the data warehouse repository to gather unstructured documents.

Based on that article it can be seen that the data warehouse is still an interesting and valuable system for future research and studies in IT and management area. The main challenges are to deal with structured data and to make the warehouse more flexible. As it was written in the previous part of this article, the goal is to use agile methodologies and combine the data warehouse with the new approaches to gather unstructured data.

However, concerning unstructured documents, it can be said that the data warehouse is not the proper repository to collect unstructured and we should use different tool for this purpose. But looking into literature we can see that even W. H. Inmon has planned to enhance the data warehouse to deal with unstructured documents, which is a DW 2.0 project.

There is still a need to evaluate the definition of the data warehouse. As it was written in this article, there are several elaborations, both research papers and academic books that used several definition of the data warehouse to ensure that the system will be correctly recognized by readers.

**References**

2. Bach M., Werner A., Duszeńko A. (2012), Selecting database structure based on the analytical processing optimization criteria
The Evolution of the Data Warehouse Systems in Recent Years


The Evolution of the Data Warehouse Systems in Recent Years


**Abbreviations**

BI – Business Intelligence
DW – Data Warehouse
GDP – Gross Domestic Product
IDC – International Data Corporation
RUP – Rational Unified Process
XP – eXtreme Programming
IT – Information Technology

**The evolution of the data warehouse in recent years (Summary)**

The article shows the key issues of the data warehouse evolution in recent years. It is divided into six general parts – in the first part a theoretical background was presented to classify the most common used definitions of the data warehouse. In the second and third part the key issues of the data warehouse evolution were presented, such as treating unstructured data. The fourth part of the paper presents the analysis of the data warehouse attributes, whether they are still valid and useful. In the fifth part the paper solutions to mitigate data warehousing weaknesses were presented. Last part shows conclusions.

**Keywords**

data warehouse, business intelligence, business informatics