The Data Warehouse as a Means to Support Knowledge Management

Michael Erdmann
Institut für Angewandte Informatik und Formale Beschreibungsverfahren
University of Karlsruhe (TH)
D-76128 Karlsruhe (Germany)
e-mail: erdmann@aifb.uni-karlsruhe.de

Abstract: This paper tries to provide a new view on the currently vastly discussed and successfully employed concept of a Data Warehouse. This view presents it in the light of Knowledge Management, i.e. a Data Warehouse can serve as a storage medium for keeping the corporate memory, or at least concerning certain types of data. It helps gaining new knowledge by delivering well integrated data to analysis tools, e.g. On-Line Analytical Processing or Knowledge Discovery in Databases, and thus becomes an important part of Decision Support Systems or Executive Information Systems. In this way a Data Warehouse, storing only data, results in growth of knowledge and may lead to enhance the enterprise's success.

The paper does not claim, that a Data Warehouse is the only thing an enterprise needs to perform successful Knowledge Management.

1 Introduction

During the last months several workshops, symposia etc. dealt with a new (or not so new) topic: "Knowledge Management" (KM). The term seems to embrace several existing research areas, which are all tied together by their common application environment, namely the enterprise. Some topics gathered under the new label are workflow management, business process modelling, document management, data bases and information systems, knowledge based systems, and several methodologies to model diverse aspects relevant when dealing with knowledge —or the like— in an enterprise environment.

One key term when discussing knowledge management became the "Corporate Memory" or "Organizational Memory". This memory serves for storing the enterprise knowledge which has to be managed. Analogous to the diverse approaches summoned together as knowledge management the corporate memory also contains several kinds of information, e.g. know-how in the heads of employees; case-knowledge, such as lessons learned; atomic, raw, or low level data, such as lists of customers, suppliers, or products, which are stored in data bases; or several documents stored as natural language texts in files. [Kühn, Abecker 97] define a corporate memory as "an enterprise-internal application-independent information and assistant system [which ...] stores large amounts of data, information, and knowledge from different sources of an enterprise."

In this paper we will show how a Data Warehouse (DWh) smoothly matches this definition and thus should be considered during KM decision processes. Although the "D" in DWh suggests that only data is stored in a DWh, this data can become valuable knowledge for the enterprise by analysing the large amounts of data with Knowledge Discovery (KDD) or On-Line Analytical Processing (OLAP) mechanisms.

Because we think "knowledge managers" should be aware of some differences between data, information, and knowledge we will try to define these three terms in section 2, although we will not back up these definitions with a comprehensive philosophical discussion. The next section then, will present the fundamental principles underlying a DWh and its contribution for knowledge mining through data analyses. In section 4 the DWh is related to KM without assuming that a DWh may solve every problem arising whilst KM processes and without presenting it as *the ultimate KM system*.

2 Data, Information, and Knowledge

In this section we will present three terms widely —but often unreflectingly— used in several IT-related (and other) communities, i.e. 'data', 'information', and 'knowledge'. The definitions will be oriented according to the three dimensions of semiotics (the theory of signs), i.e. syntax, semantics, and pragmatics [Morris 71].

[Aamodt, Nygard 95] state "there is, in general, no known way to distinguish knowledge from information or data on a purely representational basis." As we see it, this is due to the fact, that any representation is restricted to using signs (e.g. ASCII-characters, bits, or handwriting), thus there simply cannot be any distinctions. It is only through relations, that signs or representations can be separated into data, information, or knowledge. Signs can be interpreted along three dimensions. (1) The relation among signs, i.e. the syntax of ’sentences’ does not relate signs to anything in the real world and thus, can be called one-dimensional. In our eyes, signs only viewed under this dimension equal data. (2) The relation between signs and their meaning, i.e. their semantics adds a second dimension to signs. It is only through this relation between sign and meaning, that data becomes information. (3) The third dimension is added by relating signs and their ’users’. If this third dimension including users, who pursue goals which require performing actions, is introduced, we call the patterns, or signs knowledge. This relationship, which is called pragmatics, defines an important characteristic of knowledge, i.e. only knowledge enables actions and decisions, performed by actors.

To illustrate these distinctions we will give an example: *What does the sign "25" mean?* Because we can only perceive the syntactical dimension it is nonsense to ask for the meaning of this **data**. After adding a relation between the sign "25" and the real world concept of "25 meters", we can assign a meaning to the given pattern; we yielded a bit of **information** but we do not know what to imply from this information. The information does not induce or suggest any actions. So, we can ask *What does this information mean to us or any other person?* Assuming, that the sign "25" is shown on the display of an instrument indicating the distance of a landing plane from the floor underneath this information must be interpreted by the pilot in an appropriate way. His **knowledge** then may lead to certain actions to successfully finish the landing manoeuvre. As we see, knowledge —on the representational level— does not differ from data, but provided a concrete context and more knowledge to interpret raw data it makes actions possible.

Transferring these semiotically motivated definitions into the area of knowledge management, there can be seen plain analogies. It does not matter whether patterns were represented in data bases, information systems, knowledge based systems, or any other (computer) systems; they are all alike, i.e. they are all represented by signs. It is only through usage of these signs, including their various roles, contexts, and users, that they become data, information, or knowledge. In [Aamodt, Nygard 95] this kind of distinction is called *frame of reference* and states who uses patterns in what way, e.g. patterns stored in an information system are used (i.e. interpreted) by human users, whereas domain models of knowledge based systems (KBS) are used by (automated) problem solving methods (PSM) for inferencing. Thus, in the first case the user of the IS knows or learns something, whereas in the latter case one could claim, the KBS contains knowledge.

According to the three semiotical dimension identified for signs, a pattern (as data) has to be interpreted to yield information, i.e. data with meaning. This interpretation requires knowledge, i.e. knowledge has to take an active role during the interpretation process. Thus, we can further distinguish knowledge from data and information. Data, information, and knowledge embody passive objects which have to be handled within knowledge management. Knowledge alone has the capability to support knowledge management actively. Knowledge, or its owners/users are the subjects capable of acting. The enabled actions can be manifold, e.g. as we have seen processing, interpreting, and understanding data and information; learning, i.e. gaining new knowledge; or any external actions, such as selling stocks, cancelling a project, or rating the credit-worthiness of customers etc.

In the next section we will present the main features of a Data Warehouse and show how the stored DWh data can support effective actions through data analyses.

3 The Data Warehouse

This section will describe some basic concepts of Data Warehouses. The term Data Warehouse (DWh) has been defined by Bill Inmon —the *father of Data Warehousing*— as follows [Inmon 96]:

> *"A Data Warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process."*

This definition reflects the main purpose, a DWh has to support. It contains data and delivers it to executives as knowledge, they can built their decisions upon. The four named adjectives characterizing a DWh distinguish DWhs as informational systems from so called operational systems.

- A DWh is *subject-oriented* because the data it contains is structured in a way reflecting the business objects of the company (e.g. products, clients, sales). Operational systems on the other hand tend to be "organized around the applications of the company" [Inmon 96], e.g. databases handling all data relevant for booking passengers for flights. This system contains several subjects; a fact which complicates data analyses. The subject-orientation, on the other hand especially supports analytical tasks (see below) and thus, the production of knowledge.

- The second aspect, the *integration*, is the main characteristic of a DWh. A DWh contains data stemming from several sources (i.e. operational systems) which are spread all over the enterprise. These heterogeneous sources have to be integrated to access data in a uniform and clear way, i.e. all data has to be represented in an integrated way. Integration means, all data that is loaded into the DWh is transformed into a unique representation, e.g. no matter how the gender of persons is represented in several operational (source) systems (e.g. male/female, m/f, 0/1, X/Y etc.), one representation is selected and all others are transformed into this unique one. Integration of heterogeneous data sources has been investigated for some time in the IT-area [Saltor et al. 93], esp. since the growing importance of the internet and its numerous information sources. Only by defining an integrated representation analytical processing in the large amounts of data stored in a DWh becomes possible.
- A DWh is a *time-variant* collection of data, i.e. it contains current data as well as historic data. Due to that analytical processing can be done along the time dimension, thus trends and developments can be identified concerning the subjects of the enterprise, e.g. the development of sales of several products in several regions may be compared for the last twelve months. In contrast, operational systems only contain up-to-date data, thus no trends are recognizable within such a system. The DWh contains a sequence of snapshots taken periodically from operational level data.
- *Nonvolatility* of a DWh means, everything put into a DWh remains there in one way or another. Operational systems are highly volatile, i.e. records are frequently added, accessed, updated, or deleted. These read and write accesses require special mechanisms to prevent deadlocks, to prevent loss of information, and to ensure consistency. A DWh is essentially accessed read-only with the exception of loading new data into the DWh by taking snapshots at well defined points in time. This read-only access is due to the purpose to support analytical needs in "management’s decision-making".

A DWh is organized within at least two orthogonal dimensions, a dimension of time (see above) and a granularity dimension. Data loaded into the DWh from an operational system enters as up-to-date, detailed data (see figure 1). All detailed data can be aggregated under several criteria to yield lightly summarized data. These summaries can further be aggregated to yield highly summarized data, etc. E.g. daily sales could be stored at the detailed level (i.e. one snapshot of sales data is taken each day), the lightly summarized data represents weekly and the highly summarized data represents monthly aggregation. Thus several levels of granularity are stored in a DWh, although this produces some redundancy. Because of the enormous amounts of data stored in a DWh some analytical tasks only are computable within an acceptable time, if some required data is pre-aggregated. Since all data remains in the DWh it ages with time and simultaneously its importance and the chances of accessing decrease. The time horizon for a DWh (normally 5 to 10 years) is significantly longer than that for operational systems (normally 60 to 90 days) [Inmon 96]. Despite the data’s age it actually *may* be accessed in the future, so it stays in the DWh but moves to external (slower but cheaper) storage media, e.g. optical disks, tapes, or micro fiches, while the more interesting data is stored on direct access storage devices, e.g. hard disk. Even data stored in these external media is considered part of the DWh, because these data can be accessed for analyses, if needed.

Besides raw and aggregated data a DWh contains metadata describing its contents, the sources of data, and the transformation procedures converting raw data into aggregated

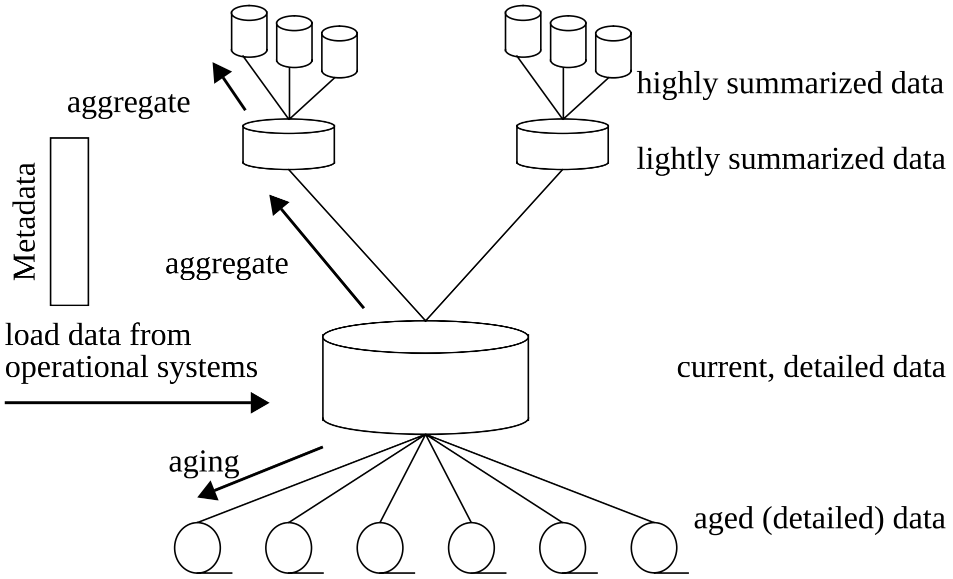

Figure 1 Structure of a Data Warehouse

data or source data into integrated, cleansed data. Metadata also serves as a navigation aid for the DWh-users, i.e. the data analysts. The analysts will consult metadata when planning data analyses.

The DWh has been defined as a "collection of data" with the goal to support "decision making processes. Essentially the DWh contains several kinds of data which are accessed through analysis front ends, such as OLAP tools or KDD workbenches, i.e. the DWh provides data for analyses which then support decision making. The possibilities provided by data analyses will be presented in the next section as one contribution of Data Warehousing to knowledge management.

4 Data Warehouse and Knowledge Management

After stating what a DWh looks like, we will point out in which way the DWh could contribute to a company wide knowledge management. In fact, a DWh could serve as one main component in a knowledge management system. The data contained in a DWh represents a large part of a company's knowledge, e.g. the company’s clients and their demographic attributes. The DWh represents an enterprise wide data collection, which is central and defines a common basis for several enterprise units accessing it. From the stored data new knowledge can be derived using technologies such as On-Line Analytical Processing (OLAP) or Knowledge Discovery in Databases (KDD).

Data analyses may consist of several reporting and visualisation mechanisms of the data, presented on different levels of aggregation, from different angles (i.e. dimensions), and using different graphical types of diagrams. These reporting facilities can be exploited interactively using OLAP-technology. Through OLAP the data analyst is enabled to formulate queries and to decide on further queries depending on the outcome of his former queries. In this way, the analyst wanders through the DWh collecting information, which he presents to the management. Recalling the definitions of data, information, and knowledge, we can recognize a similar schema. Data is stored in the DWh. The data analyst interprets parts of the data, which is represented in a way more adequate for human users. The process of interpreting data needs some knowledge and if the yielded information leads to decisions or actions performed by the management this information becomes knowledge.

Another way of gaining knowledge out of the DWh's data are algorithms provided by Knowledge Discovery in Databases (KDD). These mostly mathematical and statistical methods are able to detect knowledge previously unknown to the owners of the data. [Fayyad et al. 96] define KDD as follows:

> *"Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."*

To be able to gain valid and useful patterns out of data, it is necessary for the underlying database to contain as less noise as possible. A DWh guarantees —through its integration mechanisms— that all data is correct, so that data mining algorithms will work properly. On the solutions produced by these algorithms the management may found its decisions upon.

These technologies —KDD and OLAP— represent core mechanisms exploited within Decision Support Systems (DSS) or Management and Executive Information Systems (MIS, EIS). It is through these systems, that managements decisions are based on assured, enterprise wide, real data.

Data analyses influence or yet enable management's decisions. As an example we will present a clothing manufacturer who employs a DWh, basing on an SQL database and tools for perform OLAP. The manufacturer provides several stores across the USA with clothes in different models, in several sizes, and several colours. The first success story of the employment of the DWh and the OLAP tools relates to the number and designs of clothes delivered to the stores. The company's goal is to avoid to deliver too less as well as too much units of clothes with specific designs, sizes, and colours to each individual store, because both would result in an decreasing income, because some clothes cannot be sold, and some which could be sold, were not in stock. After establishing the DWh, a simple OLAP analysis delivered that clothes of a certain colour are sold ten times more often in Miami than in New York. Before using a DWh no queries of this type could be asked, so that a turnover of at least 30% has been lost. After gaining this new knowledge the company now can better fulfil their goals.

5 Conclusion

Successful knowledge management needs to integrate data bases, information systems, and knowledge based systems. As we have presented, a DWh can connect these three kinds of systems. It provides a wide basis of integrated data; this data can be presented via Management or Executive Information Systems (MIS, EIS). It could be interpreted as knowledge if analysis algorithms discover currently unknown patterns in the large amounts of DWh data. Newly derived knowledge or visualized information may be incorporated into the management's decision making process.

The DWh and several other more technical points —naturally discussed in the CS and AI communities— only represent one aspect of knowledge management. [Sierhuis, Clancy

97] write "knowledge management is not just about modelling problem solving and expert knowledge [or the like]. Knowledge Modelling is also about modelling the dynamics, social and cognitive, of a human activity system", i.e. the people in an enterprise must not be forgotten. In KM they play the central role as *carriers* of knowledge.

Concerning the DWh, this means that the DWh must be complemented by several other technologies and ways of working to yield successful knowledge management, i.e. a DWh is not *the ultimate solution*. Yet, there seems not to exist such an ultimate approach to knowledge management due to its immense wideness.

6 References

[AAAI 97] **AAAI-97 Spring Symposium**: *Artificial Intelligence in Knowledge Management*. Working Notes. Stanford University, Stanford, California, March 1997.

[Aamodt, Nygard 95] **A. Aamodt, M. Nygard:** Different Roles and Mutual Dependencies of Data, Information, and Knowledge - An AI Perspective on their Integration. in: *Data & Knowledge Engineering* 16 (1995). Elsevier, North-Holland1995. pp.191-222

[Abecker et al. 97] **A. Abecker, A. Bernardi, K. Hinkelmann, O. Kühn, M. Sintek**: *Towards a Well-Founded Technology for Organizational Memories*. in [AAAI 97] pp 1-7

[Chaudhuri, Dayal 97] **S. Chaudhuri, U. Dayal**: An Overview of Data Warehousing and OLAP Technology. in: *SIGMOD Record*, Vol. 26, No. 1, March 97. pp. 65-74

[Fayyad et al. 96] **U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurasamy**: *Advances in Knowledge Discovery and Data Mining.* MIT Press, Cambridge 1996.

[Inmon 96] **W.H. Inmon**: *Building the Data Warehouse*. 2nd edition. John Wiley & Sons, Inc., New York 1996.

[Inmon et al. 97] **W.H. Inmon, J.D. Welch, K.L. Glassey**: *Managing the Data Warehouse*. John Wiley & Sons, New York 1997.

[Inmon, Hackathorn 94] **W.H. Inmon, R.D. Hackathorn**: *Using the Data Warehouse*. John Wiley & Sons, New York 1994.

[Kühn, Abecker 97] **O. Kühn, A. Abecker**: Corporate Memories for Knowledge Management in Industrial Practice: Prospects and Challenges. submitted to: *Journal of Universal Computer Science (JUCS),* 1997.

[Morris 71] **Ch. W. Morris**: *Writings on the General Theory of Sign*. Den Haag 1971

[Saltor et al. 93] **F. Saltor, M.G. Castellanos, M. Garcia-Solaco**: Overcoming Schematic Discrepancies in Interoperable Databases. in: **D.K. Hsiao, E.J. Neuhold, R. Sacks-Davis (eds.)**: *Proceedings of the IFIP WG2.6 Database Semantics Conference on Interoperable Databases*. Elsevier, North-Holland, 1993. S.191-205.

[Sierhuis, Clancy 97] **M. Sierhuis, W.J. Clancey**: Knowledge, Practice, Activities, and People. in [AAAI 97] pp. 142-148

[Wu, Buchmann 96] **M.-C. Wu, A.P. Buchmann**: Research Issues in Data Warehousing. in: *Proceedings of the BTW* 1996.