JOINT ADVANCED STUDENT SCHOOL 2008, ST. PETERSBURG

KNOWLEDGE AND INTELLIGENT TECHNOLOGY FOR BUSINESS PROCESSES OPTIMIZATION (SAP)

Final Report by

Constantine Sonkin

address:
28, Grazhdansky pr., 195220
St. Petersburg, Russia,
Tel.: +7 (812) 545 4171

Institute of
Technical Cybernetics
St. Petersburg State Polytechnical University
Univ.- Prof. Dr. Vyacheslav Shkodirev

Supervisor: Prof. Dr. Vyacheslav Shkodirev
Submitted: 03.04.2008

Contents

1 Introduction

The ultimate goal of advanced data analysis is to make decisions that lead as directly as possible to benefits. Myths are common concerning data warehousing, data mining and decision support techniques. Many enterprises sometimes take unnecessary risks by trying to apply these techniques too soon (i.e., prior to understanding, cleaning, unifying information sources; defining business goals; and educating users) or by thinking that these techniques could be set loose on their data store and automatically uncover knowledge gold mines.
The process leading from data to decisions is a rigorous path involving three main IT disciplines: database and infrastructure technology, data mining, and Business Intelligence (BI).

Historically, BI systems have evolved through three main phases (see Figure 1.1):

- Phase 1. In the first phase, enterprises started to put in place structured data stores filing data relevant to their needs. In the 1980s, decision support tasks were performed centrally, with highly skilled individuals analyzing mainframe-resident data. The results were delivered to management as hard-copy reports and graphs.
- Phase 2. Competitive factors pushed enterprises toward a better leverage of the data and they adopted Decision Support Systems (DSSs) augmented, offline, by data analysis techniques such as statistics, creating a focused store of data with subject matter from at most two or three operational sources.
- Phase 3. Finally, to cope with the multiplicity and diversity of the data stores, enterprises are starting to unify and rationalize these through data warehouse frameworks. In addition to this effort and due to an increased competitive pressure, advanced data analysis techniques are also implemented to leverage further — and gain business advantages — the resulting and ever growing amount of data.

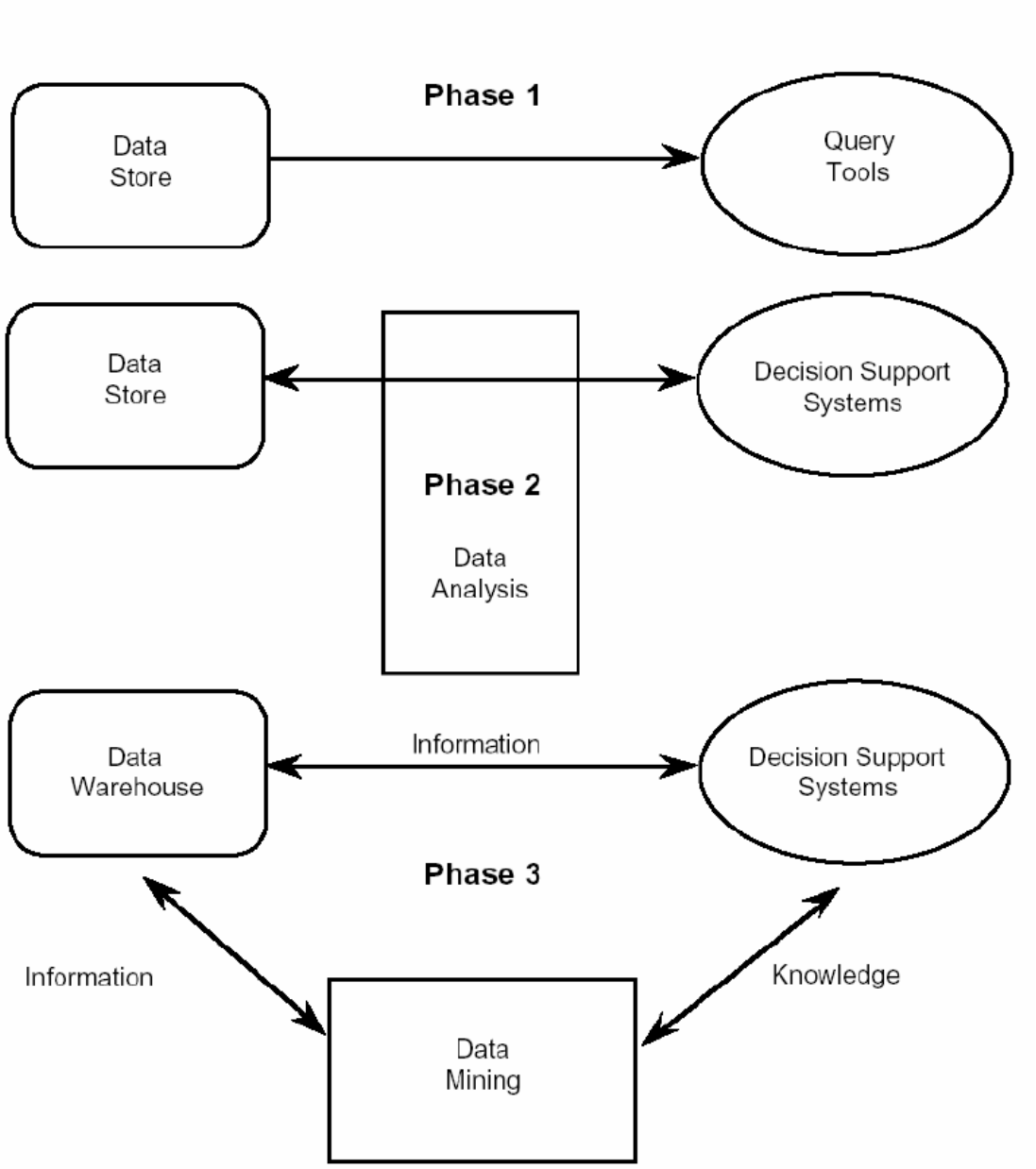

Figure 1.1: Development of BI systems (Gartner Group)

2 Data Warehousing and Data Mining

Data warehousing technology and architectures are becoming a mainstream activity, with mission-critical implementations in some early adopter sites. However, many enterprises face the high cost of implementation and experience difficulty in quantifying benefits and the return on investment. In addition, the vast majority of effort to build and manage a data warehouse still requires in-house customization of products, thus consuming significant resources.

2.1 Data Warehousing

A data warehouse is an architecture, and enterprises that focus on a single product for implementing a data warehouse increase the risk of failure. A data warehouse is more than a single product and requires significant planning of five essential components:

- Operational data source
- Data conversion and extraction
- Data warehouse database management system (DBMS)
- Data warehouse administration
- Business Intelligence tools

Operational Data Source: Data administration must take an active role to help plan extracts and to work with the business administrator to define the data needs. The data administrator can help gather the information about the operational data and assist in designing the data model that will be used for the data warehouse DBMS.

Data Conversion and Extraction: A data warehouse architecture should include extracts of operational data that are “frozen views of information” trapped in time capsules, which in some cases have some level of summarization and history associated with the view of information. Applications should provide the capabilities to perform the

complex task of integrating data from multiple sources to create a consolidated view of the data, as well as the transformation of data for use by BI applications.

Data Warehouse DBMS: The RDBMS vendors (e.g., Oracle, IBM, and Microsoft) have significantly increased the amount of research and development to improve support for data warehousing and complex DSS applications. This investment is geared toward providing strong support of complex database schemas with databases approaching several hundred terabytes.

Data Warehouse Administration: Data warehousing brings many complex administration issues that are much different from handling transactional applications and "stovepipe" DSSs of the past. New administration requirements need extensive planning to provide data usage auditing, business data model, directory management, summary tables, security, request control, query catalog and subscription services, and managing of operational data extracts.

Business Intelligence: BI empowers enterprises with systems that facilitate the access and analysis of data contained in the data warehouse. A data warehouse in combination with the right BI tools can be an important part of supporting a business mission. The selection of BI tools needs to be done after an analysis of the business and enterprise needs have been performed.

Structure of DW in broad sense can be summarized at the fig 2.1.

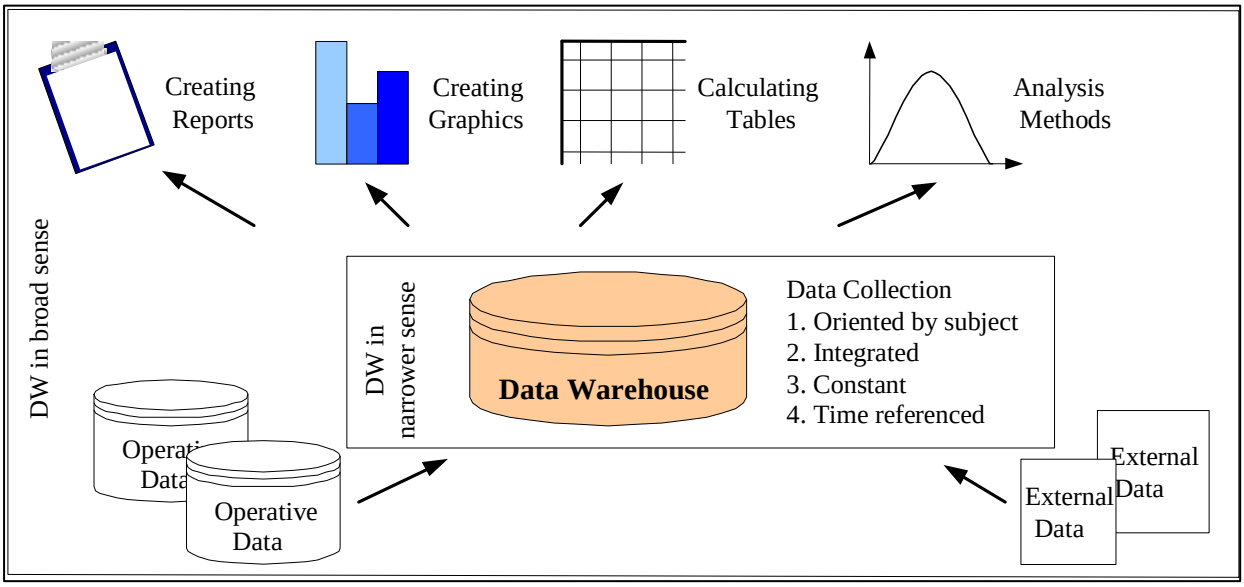

Figure 2.1: Data Warehouse in broad sense

Implementation difficulties

With most market size estimates reaching as high as $6 billion, it is clear that data warehousing has captured the imagination of the masses. But very important to keep n mind some crucial properties of DW systems, such as:

- DW can't be taken from a shelf – has to be compiled using a variety of components
- Clear idea of goals and benefits is a necessity
- DW is an architecture, not a product
- Can't be bought – need to be built

2.2 Data Mining

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies and statistical and mathematical techniques. The mining of databases is a horizontal application — that is, it is not specific to any industry. The common ingredients are a number of structured data sets and the willingness to explore the possibility of hidden knowledge that resides in the data.

Three main reasons motivate enterprises engaged in data mining activities:

- Correcting data – correction of incompleteness and contradictory information
- Discovering knowledge – to determine hidden relationships, patterns and correlations from data
- Visualizing data – to humanize the mass of data, display the data

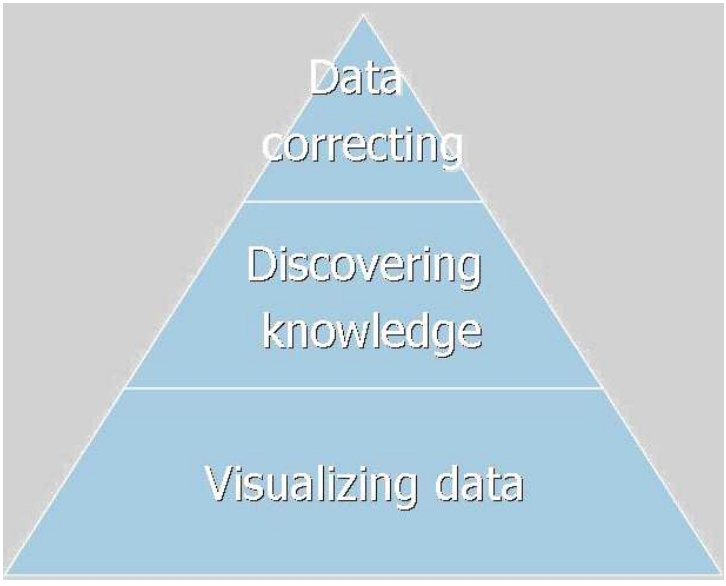

Figure 2.2: Reasons for data mining

2.2.1 Methodology

It is important to proceed with a methodology while developing data mining solutions. The main steps of the methodology are:

1. Database selection and preparation
    - identification of databases and factors to be explored
    - preparation includes filling in missing values and rectifying errors
2. Cluster and feature analysis
    - database groups are divided using clustering techniques
    - more detailed feature analysis to find the main factors
3. Tool selection
    - Many tools are available, but most are not complete and often may have to be combined with techniques or systems already developed within the enterprise. Buyers should make a thorough analysis before acquiring a tool or a technology by asking themselves the following questions:
        - How many examples can be handled at one time?
        - How much processing is required?
        - What are outputs?
        - Simplicity of update?
4. Hypothesis testing and knowledge discovery. This is the step most often associated with the term data mining.
    - hypotheses are formed and tested (top down)
    - new relationships are discovered (bottom up)
    - what-if analyses may be performed
5. Knowledge application
    - tested rules created from the discovery process can be directly added to either procedural code or into a knowledge-based system.

2.2.2 Main approaches to data mining

Data mining is the process of discovering and the main concept of solving that type of problems. A number of underlying techniques can be applied to the various functions of the data mining effort. Visualization, statistics, induction and neural networks are the most popular.

Visualization: Visualization relies heavily on the human aspect of analysis. Even the best set of rules or tables of data may reveal more information when visualized with color, relief or texture in two-dimensional (2-D), 3-D or 4-D (3-D with animation) representations. Visualization techniques may be used throughout the data exploration process, in combination with other techniques, such as induction, where it might show the number of rules generated as a result of certain parameter settings.

Disadvantages of visualization are the difficulty of representation of relationships among more than four variables and time series.

Statistics: Statistics are the most mainstream of techniques applied to data mining problems. Statistics are used universally by traditional and advanced technology researchers to perform many functions, such as clustering, factor analysis and prediction. Frequently used in combination with other technologies, statistics are often the technique of choice for initial analysis to identify predictive factors. New statistical techniques are constantly being developed, and this wide field continues to play a major role in data analysis. However, statistics often require up-front assumptions and are difficult for nonstatisticians to apply and interpret. In addition, statistics are difficult to use with large numbers of factors, particularly when nonlinearities are involved.

Induction: Induction is the process of reasoning from specific facts to reach a hypothesis. The opposite of induction is deduction, which is reasoning from a hypothesis and trying to prove it by specific facts. The facts in data mining applications are database records, and the hypothesis is usually a decision tree that attempts to divide data in a meaningful way. The decision tree can be used to create generalities (rules), with the nodes serving the decision points.

Neural Networks: Neural networks are multilayered network architectures that “learn” how to solve a problem based on examples. The two broad categories of neural networks are unsupervised and supervised. Unsupervised networks are used initially to divide data into groups or clusters according to rules set by the developer. Supervised algorithms are

used to create predictive models that capture the nonlinear interactions between factors. Both types can evaluate large numbers of factors and are able to ignore variables that do not provide value.

Neural networks are used in a wide variety of applications. They have been used in all facets of business from detecting the fraudulent use of credit cards and credit risk prediction to increasing the hit rate of targeted mailings. They also have a long history of application in other areas such as the military for the automated driving of an unmanned vehicle to biological simulations such as learning the correct pronunciation of English words from written text.

On the negative side, neural networks require extensive data preprocessing. Only numeric input is possible, and the ranges of the variables must be carefully scaled. Neural networks require numerous parameter settings, some as basic as network structure and size, and these decisions can have a strong influence on the results. Another problem is that neural network is a black box approach and very difficult to interpret.

Data mining techniques summary is shown at the figure 2.2

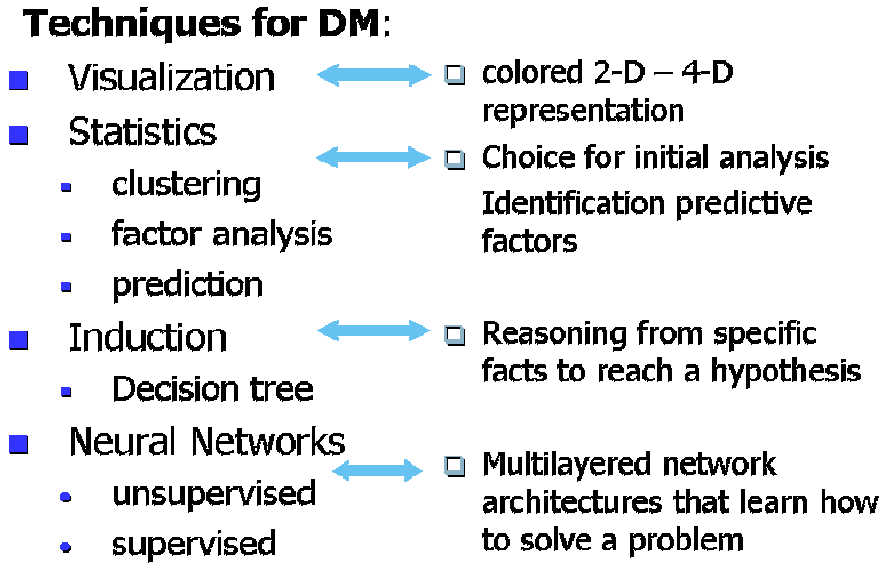

Figure 2.3: Techniques for DM

3 Business Intelligence and SAP Implementation

3.1 Business Intelligence processes

To make sure the IT environment fully addresses an organization's needs at each stage of analytical competition, companies must incorporate analytics and other business intelligence technologies into their overall IT architecture. Technologists use the term "business intelligence" (often shortened as BI) to encompass analytics as well as the processes and technologies used for collecting, managing, and reporting decision-oriented data. The business intelligence architecture (a subset of the overall IT architecture) is an umbrella term for an enterprise-wide set of systems, applications, and governance processes that enable sophisticated analytics, by allowing data, content and analyses to flow to those who need them, when they need them.

The BI architecture must be able to quickly provide users with reliable, accurate information and help them make decisions of widely varying complexity. Responsibility for getting the data that provides a clear picture of the business, major trends, risks and opportunities as well as technology and processes right is the job of the IT architect. This executive (working closely with the CIO) must determine how the components of the IT infrastructure as a whole (hardware, software and networks) will work together to provide the data, technology and support needed by the business.

Breaking the business intelligence architecture into its six elements can help IT executives leverage the analytical power of their IT investment.

1. Data management that defines how the right data is acquired and managed.
2. Transformation tools and processes that describe how the data is extracted, cleaned, transmitted and loaded to "populate" databases.
3. Repositories that organize data and metadata (information about the data) and store it for use.

4. Applications and other software tools used for analysis.

5. Presentation tools and applications that address how information workers and non-IT analysts will access, display, visualize and manipulate data.

6. Operational processes that address how important administrative activities such as security, error handling, “auditability,” archiving and privacy are resolved.

BI processes and tasks can be summarized as follows (fig. 3.1):

• Understand the business problem to be addressed
• Design the warehouse
• Learn how to extract source data and transform it for the warehouse
• Implement extract-transform-load (ETL) processes
• Load the warehouse, usually on a scheduled basis
• Connect users and provide them with tools
• Provide users a way to find the data of interest in the warehouse
• Leverage the data (use it) to provide information and business knowledge
• Administer all these processes
• Document all this information in meta-data

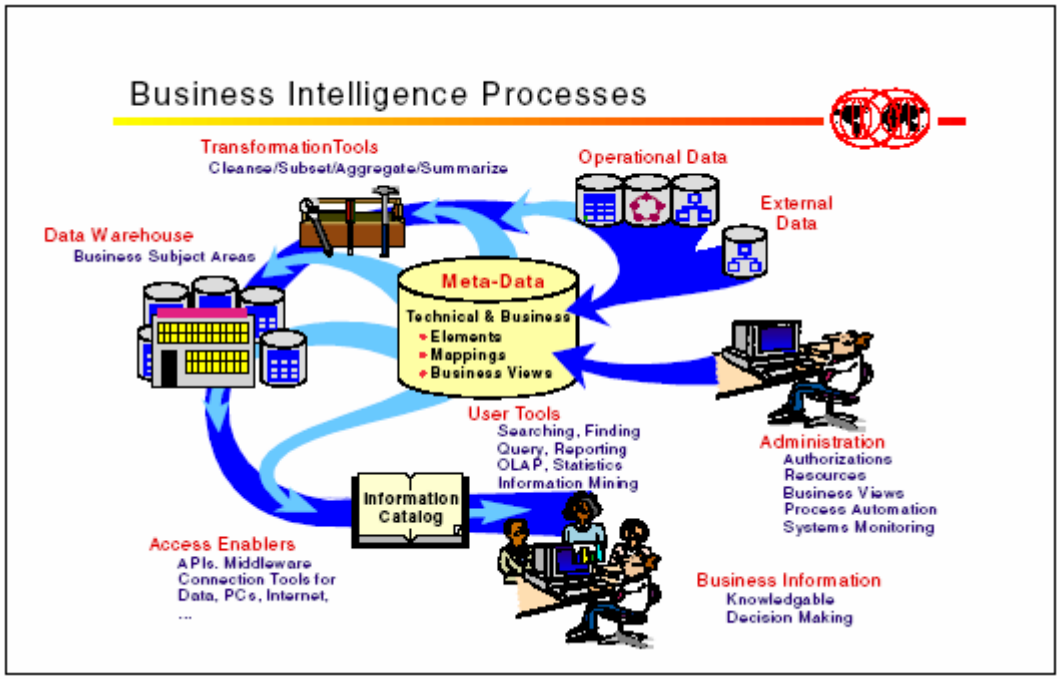

Figure 3.1: Business Intelligence Processes

BI processes extract the appropriate data from operational systems. Data is then cleansed, transformed and structured for decision making. Then the data is loaded in a data warehouse and/or subsets are loaded into data mart(s) and made available to advanced analytical tools or applications for multidimentional analysis or data mining. Meta-data captures the detail about the transformation and origins of data and must be captured to ensure it is properly used in the decision-making process. It must be sourced and maintained across a heterogeneous environment and end users must have access to it.

3.2 SAP and the Information Factory

3.2.1 SAP Development

1. The first introduction of SAP to most corporations is as an ERP vendor. ERP technology allows organizations to get a handle on the integration and modernization of daily transactions that run the operations of the business.

There are several reasons why basic transaction data is not enough to run the business:

- Not all data is incorporated in the ERP installation.
- Basic transaction data needs to be summarized, analyzed, aggregated in order for management to be able to see beyond the detail

2. SAP's first response to the need for information beyond that that is available at the transaction level was to supply the customers with a form of multidimensional technology - InfoCubes.

The overarching goals of multi-dimensional models are:

- To present information to the business analyst in a way that corresponds to his normal understanding of his business
- To offer the basis for a physical implementation that the software recognizes (the OLAP engine), thus allowing a program to easily access the data required.

The most popular physical implementation of multi-dimensional models on relational database system-based data warehouses is the Star schema implementation. BI uses the Star schema approach and extends it to support integration within the data warehouse, to offer easy handling and allow high performance solutions.

3. The next step SAP took was to initiate a data warehouse in support of the informational needs of its customer base. The evolution to an architected environment is measured by the different releases of SAP. A brief history of the SAP BW release pattern looks like:

BW 1.2b - introduction of InfoCubes and Business Content

BW 2.0b - introduction of ODS

BW 2.1c - analytical components

BW 3.0 - further enhancement of ODS into a data warehouse, along with the creation of analytical applications, partnerships, and so forth.

The architectural differences from one release to the next are very significant. One very important and positive architectural component found in the newer releases of the SAP BW product is that of the ODS which is now to a point that it fulfills the role of a data warehouse, addressing many problems that arose when there were only InfoCubes.

3.2.2 SAP highlights

The combination of SAP R/3, other mySAP components, and SAP BW, along with the analytical capabilities currently existing make the SAP support of the corporate information factory very robust.

- SAP support of ERP. This is the easiest and most natural component of the corporate information factory to be supported by SAP. In fact SAP was the worlds pioneer in this arena and there is no question of SAP's support here. The ERP foundation gives SAP a basis for gathering and managing data. This advantage is the equivalent to the being the head of the food chain. Once the data comes under the management of SAP, it is easy and natural to integrate the data and migrate the data to other parts of the corporate information factory.
- SAP support of data warehouse. The first ODS that appeared in early releases of SAP was very much like a true ODS. But each new release adds on to the ODS so

that in later releases the ODS starts to take on the characteristics of a data warehouse. The latter release of the ODS contains granular data that can be accessed in many different fashions, a smattering of historical and data integrated from the SAP R/3 environment and elsewhere. As such the latter release ODS starts to look very much like a data warehouse. Certainly from a functionality standpoint, it serves as a warehouse. In addition the ODS operates in a mode of openness, where access to the warehouse is available in many forms, and where the ODS is operable with software and technology from partnerships such as Ascential and Tealeaf.

- SAP's support of the web environment. SAP has done a very good job of supporting the web through mySAP.com. mySAP.com has interfaces from the web to and from the corporate infrastructure.
- SAP support for data marts. SAP has its InfoCubes. This technology plays a robust role of data mart support.
- SAP support for DSS applications. While other areas of SAP are strong, this area is perhaps the strongest and most sophisticated. SAP supports the concept of the "cockpit" here. The cockpit approach is very appealing to management who has the need for up to date information and a wide variety of information. SAP's offering of SEM – Strategic Enterprises Management - provides SAP with superior support here. SAP has divided its SEM into five components:
  - SEM-BIC - the component for business information collection,
  - SEM-BPS - the component for planning and simulation,
  - SEM-BCS - the component for business consolidation,
  - SEM-CPM - the component for corporate performance monitoring,
  - SEM-SRM - the component for stakeholder relationship management.

- SAP support for metadata, in particular distributed metadata.

  In order for the corporate information factory to work seamlessly, in order for the corporate information factory to become a cohesive whole, not a collection of independently operating devices, there needs to be distributed metadata. Metadata needs to be available and useful at each architectural component and metadata needs to be able to be transmitted across each component to the next component. In doing

so, metadata forms the glue that holds the corporate information factory together. SAP uses a combination of XML for the transport of metadata across components and industry conventions to decipher the metadata once the metadata has arrived at its destination and is freed from XML. In doing so SAP has tied the different architectural components together in a constructive manner. This approach promises to have long-term implications for the users of the corporate information factory which are very beneficial. It means that SAP has constructed their version of the corporate information factory for the long term, and this can only be satisfying to the long-term client of SAP.

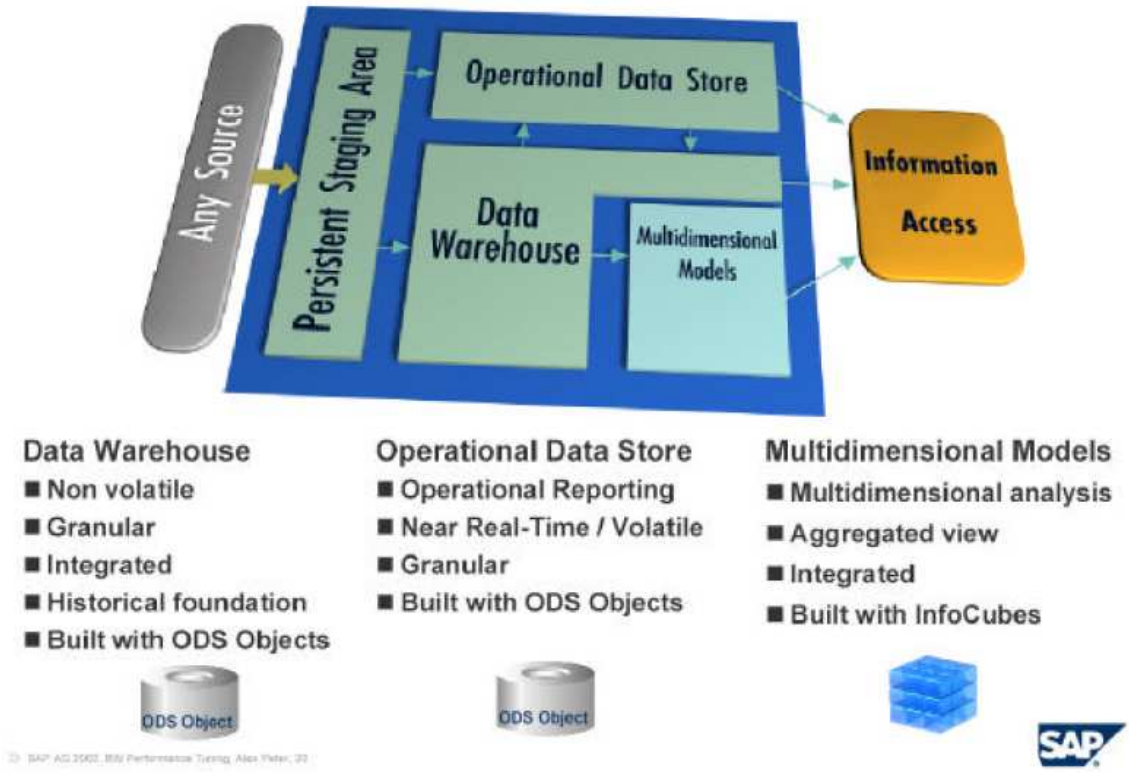

Figure 3.2: SAP Data Reorganization

4 Summary

In modern businesses, increasing standards, automation, and technologies have led to vast amounts of data becoming available. Data warehouse technologies have set up repositories to store these data. Improved Extract, transform, load (ETL) have increased the speed of collecting the data. OLAP reporting technologies have allowed faster generation of new reports which analyze the data. Business intelligence has now become the art of sifting through large amounts of data, extracting pertinent information, and turning that information into knowledge from which actions can be taken.

Business intelligence software incorporates the ability to mine data, analyze, and report. Some modern BI software allows users to cross-analyze and perform deep data research rapidly for better analysis of sales or performance on an individual, department, or company level. In modern applications of business intelligence software, managers are able to quickly compile reports from data for forecasting, analysis, and business decision-making.

In 1989 Howard Dresner, later a Gartner Group analyst, popularized BI as an umbrella term to describe a set of concepts and methods to improve business decision-making by using fact-based decision support systems. Nowadays this technology is the important part of successful business process.

Bibliography

1. Basics - Reporting & Analysis - Modeling & Staging - Data Mining, Helmut Krcmar SAP UCC at Technische Universität München
2. Business Intelligence, Gartner Group, http://www.gartner.com
3. Data warehouse, Wikipedia, http://en.wikipedia.org/wiki/Data_warehouse
4. An Introduction to Data Mining, Kurt Thearling, www.thearling.com/SAP
5. Data mining, Wikipedia, http://en.wikipedia.org/wiki/Data_mining
6. Building Data Mining Applications for CRM, Alex Berson, Stephen Smith, Kurt Thearling
7. The New Science of Winning Harvard Business School Press, March 2007
8. The Architecture of Business Intelligence, Thomas H. Davenport and Jeanne G. Harris
9. Business Intelligence Architecture on S/390, ibm.com/redbooks
10. Multi-Dimensional Modeling with BI, SAP, May 2006
11. BI Data Modeling and Frontend Design, SAP Community Network https://www.sdn.sap.com/irj/sdn/
12. Business intelligence, Wikipedia,

    http://en.wikipedia.org/wiki/Business_intelligence