BABEŞ–BOLYAI UNIVERSITY, FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION
DOMAIN: CYBERNETICS AND STATISTICS

Doctoral Thesis Summary

Design and Implementation of Data Warehouses for Business Intelligence applied in Business

SCIENTIFIC ADVISOR
PROF. UNIV. NIŢCHI ŞTEFAN IOAN, PHD

PHD CANDIDATE
NAGY ILONA MARIANA

Cluj-Napoca
2012

Table of Contents

Extended Table of Contents

Key Words

Business Intelligence, data warehousing, software development methodologies, metadata, data warehouse architecture, implementation framework, prototype, process automation;

Introduction

Advances in the information technology[1] domain have guided the evolution of data processing systems from the early stages of single support and stand-alone applications to the comprehensive Business Intelligence and analytical systems of today's informational environment. Within this broad context, data warehousing defines an extensive blend of technologies emerged in early 1990s as result of advances in the area of data processing capabilities achieved in computer-based information systems.

Data warehousing represents a component of the overall Business Intelligence framework, which covers an ample range of applications and tools that help analyze large amounts of data and transform it into understandable information and business knowledge. It is designed to handle an informational environment in which a series of components enable the gathering and integration of data from across the enterprise in order to provide business users with consolidated, structured data and improve the decision-making processes.

The repository of this comprehensive data warehousing technology, defining the storage environment, is known as the data warehouse. The data warehouse represents a model of enterprise data, especially structured for facilitating querying and analysis processes on integrated and consolidated data. It is an essential and dominant component of data-driven decision support systems and its main goal is to enable business users to make effective tactical and strategic decisions based on factual data, by answering business questions timely and accurately. For achieving its goal, the data warehouse is defined by particular data models that specify the structure of data in the repository. These data models, optimized for querying and analysis, are created in a stable, consistent and predictable manner through different data modelling techniques. Querying and analysis are enabled by means of various types of metadata meant to describe the structure of an organisation's use of information and to attach semantics to the business processes and the resulting data.

Considering its complexity level, data warehousing solutions development demands a structured and planned approach, defined in the form of a methodology, as well as an appropriate architectural framework. Methodologies are designed to achieve predictable results according to well-define requirements and to provide repeatable, trainable and consistent development processes. Architectures represent the structures that bring all the components of the data warehouse together and provide a solid basis for enterprise-wide data integration. The selection of a suitable methodology and architecture determine the overall success of the data warehouse solution implementation. Another essential aspect in the data warehouse development regards the employment of a framework able to provide a set of guidelines for describing its components and their interoperability and to support the existence of an environment of reusable processes, integration, consistency and flexibility in information delivery.

[1] As defined by the Information Technology Association of America (ITAA), *information technology* is "the study, design, development, implementation, support or management of computer-based information systems, particularly software applications and computer hardware".

i. Problem Statement And Research Goals

As a comprehensive component of the informational environment, the data warehouse is defined by the concepts of data model, development methodology, architecture, and framework. Its design and implementation represent a challenging endeavour exposed, similarly to other complex projects, to higher risks of failure, as explained below.

The literature and especially practitioners' reports present many cases in which the success rates of data warehousing development are negatively influenced by the high costs and the extent of time required for the planning, design and implementation activities. According to Adelman et al. [6], three of ten situations that generally account for data warehouse development failure are determined by the following reasons: 1) the project is over the allocated budget; 2) the delivery schedule is exceeded, and 3) some project expenses are not justified. Other acknowledged risks include frequently changing requirements from behalf of business users, deficient project management activities, poorly architected solution, and lack of data quality, etc. Beneficiary enterprises underestimating the total costs of data warehousing implementation and companies that undertake the actual solution development are highly vulnerable to these risks, as activities such as design, implementation, and maintenance, etc. involve considerable financial effort and are generally seen as lengthy and laborious.

Nevertheless, data warehousing development is guided by methodological and architectural approaches aimed at facilitating the delivery of successful solutions within project-defined boundaries. In this case, the literature and numerous best practices provide comprehensive guidelines for data warehousing design and implementation, which enterprises may use and adapt to their specific needs. These guidelines focus mostly on project management activities, data warehouse models and modelling techniques, and reference architectures. However, optimizations of the implementation process, for instance of highly repetitive and time-consuming tasks, have not been yet sufficiently covered in the literature. We consider that by proposing a framework for handling these tasks under well-defined conditions, the overall costs of data warehousing implementations may be reduced significantly.

Thus, our main research goals are established by the cost reduction (i.e. of the overall data warehousing solution development) and implementation efficiency requirements. In order to achieve these goals, we set out to make a comprehensive presentation of the theoretical aspects concerning the previously introduced concepts, namely various decision support technologies and their importance in the business environment, methodologies and specific data modelling techniques, enterprise architectures and frameworks, etc. These concepts are essential in successful data warehousing solutions development, and cover the following perspectives: the *definition* of logical and physical data models of the data storage structures, various types of metadata, and data management activities; the solution *development methodology* (e.g. project management and planning activities, best practices, and enterprise-wide standards to which it should comply, etc.); *architectural aspects* (e.g. utilized systems, data and involved processes); the *framework* defined for providing a set of enterprise

guidelines for the development of components and their interaction in the informational environment; and the *implementation* details (e.g. specific tools and applications, implementation team, time boundaries, etc.).

More specifically, our research goals are defined for each particular concept, as follows:

- *Methodological approach*

  The methodology defines a set of principles that govern the design and implementation of software solutions. We aim to introduce in our thesis several generic methodologies and discuss their appropriateness in the specific context of data warehousing solutions. Our intention is to select a suitable methodology according to our defined project requirements, follow the provided development guidelines, and validate the obtained results.

- *Data model*

  The data model describes, from both logical and physical perspectives, the layout and properties of the data structures designed to store the data in the operational and analytical environments. Our goal, regarding the data model, is to make a thorough presentation of the data modelling options and techniques, as well as other related concepts specific to the data warehousing environment. We aim to select and enhance a dimensional modelling methodology, and use it in the development of our proposed enterprise-wide data warehouse data model.

- *Architecture*

  The architectural plan stands at the core of data warehousing solution design and implementation and is thus essential in its development process. Our goal is to introduce the most common type of architectures and their characteristics, select an appropriate architecture, according to well-define criteria and compatible with the methodological approach, and use it as foundation for our data warehousing solution development.

- *Framework*

  The framework defines the boundaries of the data warehousing system, the various components and the interaction between them. Considering this aspect, our proposal regards the development of a framework designed to achieve the automated implementation of the data warehouse model, as part of our contribution to the thesis. Our goal is driven by the requirement of time and costs reductions involved in the data warehousing solution development process.

- *Implementation*

  Regarding the implementation of the data warehousing solution, we intend to design and implement a prototype for creating specific data structures in the data warehousing environment in a partially automated manner, based on the proposed framework. The effective implementation process is achieved in the SAP Business

Warehouse environment from various types of technical metadata. The prototype represents our practical contribution to the thesis.

ii. General Outline of the Thesis

Considering the broad context to be covered, we organized our thesis into six main chapters. A beginning section is dedicated to introductory notions, while a final chapter outlines the conclusions, dissemination of results and future research directions. The *Introduction* highlights the motivation of our thesis and the research objectives determined by it. We describe the main challenges encountered in the area of data warehousing development, which dictate the improvement possibilities and thus determine our goals, introduce the structure of the thesis and summarize the focus of each chapter. Five main chapters describe fundamental aspects of Business Intelligence and data warehousing technologies, data models and data modelling techniques, data warehouse architectures and frameworks, as presented in Figure 1.

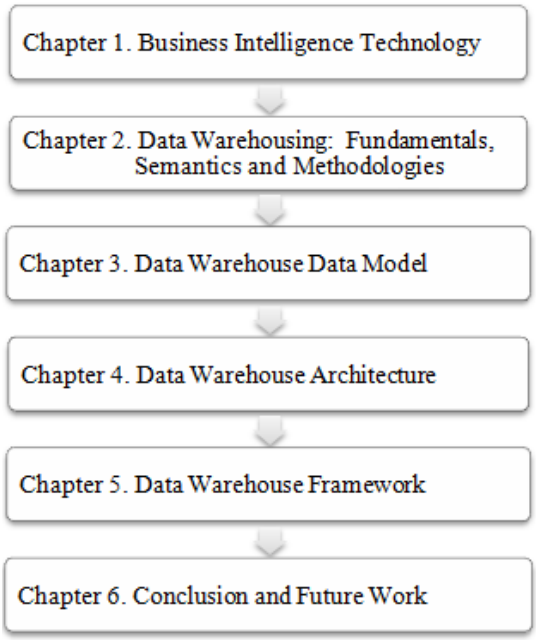

Figure 1. Thesis General Outline

Chapter 1. Business Intelligence Technology

The first chapter is dedicated to introducing the context of the business decision-making processes and the technologies that efficiently support them. Our main goal is to place the data warehousing theme in the overall Business Intelligence framework. For this reason, we outline briefly the historical use of data, examine the evolution of intelligent decision support systems and discuss their benefits and importance in the decision-making process. We also analyze the various definitions, architectures and development lifecycles of the Business Intelligence systems, the relation between the Business Intelligence and the data warehousing technologies, and the role of the data warehouse concept in the analytical framework.

In today's information society[2], the concept of information is seen as an essential component that actors of the business environment (e.g. organisations, enterprises, etc.) need to leverage in order to acquire a deeper understanding of the occurring processes, improve the decision making, and increase the reaction to change. The abundance of existing internal and external data and information may be effectively exploited in the benefits of organisations and enterprises by means of intelligent systems, such as Decision Support and Business Intelligence Systems.

Advances in the information technology field have guided the evolution of data processing systems from the early days of single support and stand-alone applications, such as Management Support Systems, to the comprehensive Business Intelligence and analytical technologies of today's informational environment.

Decision support systems represent computer-based systems that process data collected from various source systems and prepare it for analysis purposes. These systems facilitate the access to enterprise-wide data, provide processing capabilities, and enable the analysis of advantages and disadvantages of various alternatives. Managers and executives may therefore base their decisions on reliable and accurate information, and increase the quality and responsiveness of the decision-making process. Ultimately, DSS are meant to provide business intelligence in an effective and suitable manner to decision makers.

Business Intelligence systems comprise a broader category of applications and tools, from data acquisition, transformation and storage, to on-line analytical processing and interactive presentation to the end users. Their main goal is to offer a framework for improved decision-making process by supplying business people with the right information at the right time. While BI represents a comprehensive technology, DSS are smaller in scope and mostly reduced to a computer program or application; moreover, Decision Support Systems may be integrated in the Business Intelligence environment, as part of the overall BI framework. From an architectural perspective, BI establishes the data warehouse as a component of the repository element and the foundation upon which is built. These two comprehensive

[2] An exact definition of the term *information society* has not been yet universally accepted. We refer in this case to the definition provided by N. Moore: "a society where […] information is used as an economic resource, it is intensively exploited by the general public in their activities as consumers; and behind it an information sector is developed within the economy" [125].

technologies are complementary; however, they may also exist independently from one another.

The development of BI systems follows a generic cycle specific to most information systems, which includes planning and direction activities, data collection, information processing and storage, analysis and production operations, and dissemination of the produced information among intended addressees. The phases of the development cycle are intended to ensure that reliable and accurate information gathered from across the enterprise is supplied to business users upon their requirements.

The presented technologies are meant to facilitate the processing of data from multiple enterprise-wise sources into understandable and valuable information which business user may use to improve their decision making process. Regardless of their complexity, the employment of either one, according to users' needs, significantly increases responsiveness to changes occurring in the business environment.

Chapter 2. Data Warehousing: Fundamentals, Semantics and Methodologies

We dedicate the second chapter to making a thorough presentation of the data warehousing technology fundamentals, semantics and development methodologies. We define data warehousing as a blend technologies and analyze the data warehouse concept and its role as storage component in the overall analytical framework. We also present the data warehouse specific repositories, namely the data warehouse and the data marts, the main differences between them and their role in the data warehousing landscape. Moreover, we cover extensively the semantic aspects of integrated data, by presenting different types of metadata, their management and importance in the data warehousing environment. Finally, we analyze several development methodologies and discuss their appropriateness for data warehousing solutions.

Data warehousing emerged in early 1990s as a consequence of advances regarding enterprise-wide data processing capabilities achieved in the information technology field. As comprehensive technology, data warehousing handles an informational environment in which a series of components enable the gathering and integration of data from across the enterprise in order to provide business users with consolidated structured data for improved decision making.

Data warehousing represents a comprehensive technology used for handling the analytical environment of an enterprise through a series of components that enable the collection and integration of data from various sources, in order to transform enterprise data into strategic information. The collected data is consolidated and structured under a common model within this environment, and prepared for user consumption with the main goal of improving the decision making process. A data warehouse is defined as a storage component of the overall data warehousing technology, a repository of integrated, subject-oriented, non-volatile and time-variant data. Since data is collected and stored over a long period of time, various analytical and data mining tools may perform computations, mathematical and statistical calculations for producing valuable business insight, detecting trends or identify patterns. The data warehousing environment relies on storage structures modelled with different techniques (e.g. entity-relational, dimensional modelling), which improves the performance of the complex analysis performed against large amounts of data.

Performing such complex analysis requires that the meaning of the warehousing repository data be understood entirely, along with its provenance, heritage and lineage, which is achieved by means of metadata. Metadata is not only "data about data"; it has rather a broader sense and role, as it concentrates the totality of information and knowledge of the actual data existent within an enterprise. Metadata captures general and specific characteristics, offers a context and meaning to the raw data and creates the semantic layer of the organisation's informational system. This semantic layer ensures a proper interpretation and understanding of the organisation's data by all the actors involved in its exploitation and usage. The enterprise's informational systems contain various types of metadata, from

business and technical metadata, static and dynamic metadata, to descriptive, structural and administrative. In order to accomplish its functions, metadata has to be properly managed. Metadata management has an essential role in ensuring the proper functioning of the enterprise's activities, especially in the four areas where it is commonly deployed: design, operation activities, management and governance. It also helps minimizing the efforts of data warehouse administration and improving the extraction of information from the analytical environment. Within the data warehousing environment, the metadata collected from various sources is stored in the metadata repository, which aims to offer consistent and reliable access to the data and facilitates querying and browsing operations from the end-user.

Considering the complexity of the data warehousing technology, the development of such a solution requires a structured and planned approach, defined in the form of a methodology. A methodology is aimed to provide a repeatable, consistent and reliable set of steps or guidelines for achieving a predictable result (i.e. the product or software solution). Methodologies may be formal (characterized by a staged approach and a well-defined set of activities) or informal (defined in best practices, various courses, etc.); data-driven (based on the analysis of the corporate data model), goal-driven (based on the company's goals and analysis of business processes) or user-driven (focused on implementing business strategies), etc. Most software development methodologies (e.g. the *Waterfall model*, the *Incremental model*, the *Spiral approach*, the *RAD model*, the *Agile* methodologies) are suitable for a certain complexity of data warehousing development. However, two main approaches are considered reference models for data warehousing development, namely the Imnon and the Kimball methodologies. Inmon follows a spiral development and recommends the building of a large data warehousing solution – the enterprise data warehouse, while Kimball focuses on requirements oriented methodologies which facilitate the delivery of faster results and increased user satisfaction. Nevertheless, choosing the appropriate methodology depends highly enterprise's needs. Thus, following certain patterns in the selection processes increases the chances of successful implementation.

Chapter 3. Data Warehouse Data Model

In the third chapter, we cover essential aspects of the data model and the data modelling concepts in the data warehousing environment. We introduce various types of data models, their fundamental characteristics and two reference modelling techniques (i.e. Entity-Relationship and dimensional) used for the development of the data warehouse schema. We argue the importance of dimensional modelling for building analytical-specific data structures, and present a comprehensive review of the research efforts in this area. Our main objective in this chapter is to determine a suitable methodology for deriving dimensional data models from the Entity-Relationship schema of the underlying operational systems that represent the main source of data for the analytical environment. We apply a derivation methodology considered appropriate on a reinsurance data model (i.e. our case study) and enhance it with business specific modelling options. The implementation of resulted dimensional model is presented in the last chapter as part of our practical contribution to the thesis.

As repository of the data warehousing solution, the data warehouse is defined by particular representations of data and relationships between data. These representations, known as data models, are designed to ensure a complete documentation of the informational landscape in terms of existing processes, entities, relationships, and data flows, etc. Data models represent the outcome of data modelling techniques that define and analyze the requirements expressed by business users and support the enterprise's business processes.

The data warehousing environment acknowledges various types of data models and modelling techniques that stand at the basis of the specific data structures and determine the way data is stored in the repository. Two particular data models, namely the Entity-Relationship (ER) and the dimensional model, are employed in the different storage layers of the data warehousing systems. Our proposed data warehousing solution uses both these models for defining the general enterprise data warehouse schema, as explained throughout this chapter.

While the Entity-Relationship is a standardized technique applied predominantly in the operational environment, the dimensional modelling technique, specific to analytical systems, presents no generally accepted approach of obtaining the corresponding data model. Nevertheless, the literature abounds in proposals regarding dimensional modelling methods and the derivation of the dimensional model from various sources, such as user requirements, business processes or existing Entity-Relationship data models. We analyze several of these approaches, discuss their strengths and weaknesses, and formulate opinions regarding their adequacy in data warehousing development. Our main objective is to present a case study in which we derive the dimensional data model of a reinsurance business process by applying the methodology that we consider suitable. Since the enterprise's operational systems represent the main source of data for the analytical environment and also because we consider appropriate to design a solution that is able to integrate smoothly this data into the data warehousing systems, we focus our research on the transformation of the ER model into the

dimensional data model. Moreover, we enhance the selected approach with business specific modelling options. Finally, we use the resulted dimensional model in the implementation of the data warehousing solution achieved through an automation framework, which we describe as part of our practical contribution in the last chapter of our thesis.

As prerequisite activities, we define both the Entity-Relationship and the dimensional data models and modelling techniques and present extensively all the basic and advanced concepts related to them.

Data is represented, both in the operational and in the data warehousing environments, by means of diagrams, using texts and symbols meant to facilitate its understanding to the reader. These diagrams, known as data models, are obtained through various software engineering processes, among which the most commonly used are the Entity-Relationship and the dimensional modelling techniques.

Data modelling techniques are intended to define and analyze requirements expressed by business users and produce data models able to support the enterprises' business processes. The outcome of these modelling processes is represented by various types of data models, defined at different levels of abstraction (e.g. conceptual, logical, and physical). The resulted models have an essential role in describing the enterprise data and its characteristics, as well as enforcing specific business rules. They are used for enabling the management of data as a resource for the enterprise, for integrating the enterprise-wide information and defining a common architectural model for the entire informational landscape, and for designing data repositories, such as databases and data warehouses.

The ER modelling technique presents specifics adequate for modelling transaction data (e.g. normalized data model with minimum redundancy, dependencies, and inconsistencies; large number of entities; high quality data, etc), appropriate for frequent insertion, update and deletion operations, and is thus used extensively in the operational systems. Oppositely, dimensional modelling is defined by certain characteristics, such as smaller number of entities, intuitive presentation of data, analysis and querying optimized models, etc., that makes it good candidate for the analytical environment. However, both these techniques may be used for modelling data in the analytical environment: the ER technique is normally used for defining the data model of the central data warehouse repository, which comprises integrated and consolidated data stored in normalized data structures, whereas the dimensional modelling is used for modelling the data mart structures of the presentation layer.

Dimensional models are seen by many authors as restricted forms of ER models, which determines a relatively straight-forward mapping to be achieved between them. Thus, the dimensional models of the data warehousing environment may be derived directly from the ER schemas of the underlying operational systems. This approach is supported by various data-driven methodologies and suited for our data warehousing solution development case, which comprises a layer of integrated data, namely the data warehouse repository, and several data marts supplied with data from this layer. Nevertheless, considering their fairly complex nature, the literature acknowledges numerous approaches defined for producing dimensional data models as abstract representations of enterprise data. Most of them

however, due to various factors, such as increased complexity levels, novel notation systems, different graphical representations, etc. have not been applied outside the research area and the academic environment. The dimensional modelling methodologies used in the industry are generally based on informal approaches and best practices.

In order to achieve our goal of building a comprehensive data warehousing solution, we follow a well-defined top-down approach proposed by W.H. Inmon [81]. We describe in this chapter the guidelines provided by this approach for building the data warehouse layer, and apply and enhance a methodology proposed by Moody and Kortink [121] for the development of the dimensional data model. We exemplify the methodology through a case study of a reinsurance business process modelling and contribute to the development processes through enhancements regarding the business aspect of the modelling process, namely reinsurance specific modelling options (e.g. analysis of the data model's utility and integration of appropriate dimensions in the dimensional model, representation of financial data through various currencies, handling of changes in the dimensions, etc.). We also evaluate our resulted data model against a series of features that models should have in order to support advanced uses in data analysis, and conclude that it complies with most requirements and thus determines a valid representation of data.

Chapter 4. Data Warehouse Architecture

The fourth chapter is dedicated to presenting a series of data warehouse architectural aspects and main development approaches. We describe the fundamental characteristics of the most common architecture types used in the data warehousing environment, as well as two reference architectural frameworks and implementation models (i.e. the Inmon model and the Kimball model). As guideline, we present a series of factors that influence the architecture selection process, discuss the appropriate framework for every combination of these factors and outline a series of elements that determine successful implementations of data warehousing solutions.

Data warehouses are defined within the scope of the comprehensive technology that covers the enterprise's informational landscape, as repositories of data collected, integrated and consolidated from various and often heterogeneous sources. The complexity involved in the management, transformation and integration of this data, both from inside and outside the enterprise, determines the development of such solutions to be regarded as challenging attempts.

The literature, as well as numerous best practices, provides comprehensive guidelines to data warehousing design and implementation, which enterprises may use and adapt to their specific needs. Most importantly, the development process requires the selection of an architectural framework and a suitable methodological approach, which have an essential role in the overall success of the data warehousing solution. Their selection is based on various factors, including the existing informational infrastructure, the business environment, the desired control structure, the capability of the technical environment, the involvement of stakeholders and the enterprise's financial resources [13].

Due to the complexity and the broad scope of the data warehousing solution, its implementation architecture is significantly different and more complex than the classical database architecture. The data warehouse architecture aims to determine a solid foundation for the integration and consolidation of data from across the enterprise, defined by an intuitive design, and to provide an overall framework for effective development and deployment of all its components, grouped in three main areas: data acquisition, storage repository, and information delivery.

The literature acknowledges several types of data warehouse architectures, among which the most common are: the independent data marts, the data mart bus architecture, the enterprise data warehouse, the centralized and the federated architectures. The independent data marts architecture is generally implemented in small enterprises, being characterized by stand-alone departmental views of data, usually extracted from operational source systems. Although more efficient in terms of development resources, independent data marts increase data and process redundancy, have limited scalability, lack in data integration, and are unable to provide an integrated view of enterprise data. The data mart bus architecture overcomes some of these disadvantages, by providing a comprehensive framework to integrating separate

departmental views along a bus structure for an enterprise-wide view of the data. The individual data marts are developed by using conformed dimensions and considering the user requirements and the underlying business processes; while the specific structures built with dimensional modelling techniques, allow the storage of both atomic and summarized data. The enterprise data warehouse architecture represents the most complete and complex architectural type in the data warehousing environment and is the result of a comprehensive enterprise-wide analysis of data requirements. Its main goal is to provide an integrated data foundation, defined by atomic level data maintained in normalized form in the data warehouse, and to enable the building of several multidimensional views of aggregated data supplied from the underlying data warehouse. The centralized data warehouse architecture presents similar characteristics to the EDW architecture; however without including the dependent departmental views (i.e. data marts). The federated architecture consists of a set of data warehouses organized separately and scattered geographically, which operate in a semi-autonomous way, and are viewed as one large data warehouse. This architecture type is specific to large organisations which have acquired and merged other units with specific Business Intelligence solutions that have not been discarded, but rather leveraged in an integrated manner.

Regarding the implementation of these architecture types, there are several distinct approaches presented in the literature. Two main approaches, namely the top-down (achieved by the Inmon model) and the bottom-up (achieved by the Kimball model) stand out as reference architectural and methodological implementations. The main idea behind the Inmon model is the development of comprehensive enterprise data warehouse architectures: a central repository (i.e. the data warehouse) which stores enterprise-wide integrated and consolidated data, and a series of data marts offering a multidimensional view of data for effective analysis and reporting purposes. Building EDW architectures requires complex planning and design activities performed at the beginning of the project meant to resolve potential issues regarding data integration, security aspects, quality and standards, and the overall data model. This architecture offers thus consolidated data definitions and enforcement of business rules across the enterprise. Departmental dependent views developed subsequently with multidimensional modelling techniques are supplied from the central data warehouse repository. In a top-down approach, followed by the Kimball model, the data warehouse implementation is based on the creation of several dimensional views of enterprise data and their integration along a bus structure (i.e. conformed dimensions) in order to form a dimensional enterprise data warehouse view. The data marts are created based on the requirements specific to each department. Although it offers less data integration and consolidation across the enterprise, this approach is more widely accepted as it requires less implementation efforts and delivers immediate results. A third approach, known as hybrid, attempts to combine the advantages of the top-down and bottom-up methodologies by determining the degree of planning and design required for the global approach to support integration, and building the data mart structures with the bottom-up approach.

Considering the presented implementation options, the selection of the appropriate approach is not a simple assignment. The choice of an architectural framework that suits the needs of the enterprise is influence by various factors, such as information interdependence between

the enterprises' departments, urgency of project implementation, task routineness, strategic view of the data warehouse, the amount of resources available and allocated for the data warehousing solution development, the perceived ability of the in-house IT staff, etc. The combination of these factors favours the selection of a certain type of architecture. However, the simple selection process does not ensure the success of its implementation and deployment. Several elements, referring to organisational, environmental, project-related, technical and educational aspects, determine the extent to which the data warehousing solution is accepted by the end-users and is capable of supporting effectively the decision-making processes.

Chapter 5. Data Warehouse Framework

We dedicate the fifth and last chapter of the thesis to describing our practical contribution, namely the framework and prototype proposal for the automated implementation of the data warehouse schema in the analytical environment. In order to justify our proposal, we examine briefly several of the most utilized development frameworks in the academic environment and the industry, and discuss the utility of automation in data warehousing implementations. We structure our contribution in two main parts: the first part presents the architecture of the proposed framework, along with its components, detailed description, and the interaction between them; the second part describes the prototype design and implementation particularities for the SAP Business Warehouse environment. We also assess the utility of our framework and prototype proposals, as well as their importance in facilitating the development of the enterprise data warehouse, and thus in enabling the less costly creation of an intelligent system for supporting the business decision-making process.

Considering the characteristics of data warehousing solutions, discussed throughout the previous chapters, the allocation of substantial and varied resources on behalf of enterprises, as well as sustained commitment from the stakeholders, is easily understood for the elaborate and costly development process. Development activities assume the examination of the overall enterprise informational landscape; thus, the existence of a comprehensive methodology that guides the data warehousing design and implementation steps along a solid architecture is nearly mandatory. Moreover, costs, mainly reflected as financial resource investments and time-to-delivery[3] expenses from a project management perspective, and especially their reduction, represent fundamental concerns for all enterprises. The necessity of diminishing costs has led to the partial or complete automation of some process in data warehousing design and deployment, such as conceptual and logical modelling of the data warehouse schema, data extraction, transformation and loading, etc. Automation does not cover, however, all the project development phases, mostly due to the high influence of the business aspects in analytical environment. Nevertheless, we consider that automation may be extended, within certain boundaries, to the implementation phase and the creation of data structures and ETL processes. This requires, foremost, the definition of a comprehensive architectural framework, which we cover in this chapter as our main contribution to the thesis.

According to [55], a framework is defined for providing a philosophy or a set of guidelines that describe the look, feel and interoperability of software applications. We focus, thus, on defining an implementation framework concerned with the automated generation of data structures from technical metadata and begin with the definition of the problem statement for our proposal. We introduce the main characteristics of data warehouse development frameworks, enlist some of the most widely utilized frameworks in the academic and practitioners' world, and analyze the necessity of automation in the data warehousing

[3]According to the Business Dictionary, http://www.businessdictionary.com, time-to-delivery refers to the extent of time passed from the beginning of a project (product) to its delivery into production.

environment. We present the framework architecture, followed by a detailed description of each component and their interaction within the well-defined boundaries of the environment. Specific implementation details of our proposed prototype, the assessment of the automated implementation utility, as well as the importance of our contribution, are introduced in the second half of this chapter.

Building an enterprise-wide data warehousing solution represents a complex activity, which requires the use of a solid development framework and effective project planning activities. The literature acknowledges several architectural and methodological frameworks used in the data warehousing environment, each describing the different structures and processes, as well as the sequence of steps followed for the development of these comprehensive solutions. Successful design and implementation processes are supported by compatible architectures and methodologies. In our attempt to develop an enterprise-wide data warehousing solution, we adhere to Inmon's consistent architectural and methodological framework. Thus, from an architectural perspective, we follow a top-down design of a comprehensive enterprise-wide data warehouse, with a central repository of integrated and consolidated data (i.e. the data warehouse layer) and several departmental views supplied from the data warehouse layer. Considering the numerous structures and processes which define a data warehousing solution, we propose an automated implementation framework and a corresponding prototype, based on the assumption that repetitive and time-consuming development activities may be carried out more efficiently and in a shorter period of time. The implementation prototype achieves the automated creation of data warehousing storage structures for the enterprise data warehouse and the data mart layers, and corresponding ETL processes from technical metadata. Among the automation benefits, which justify our proposal and aim to reduce the implementation effort, we mention: the creation of software components that comply with well-defined syntax and constraints, and thus decrease the human error factor; the standardization of these components, which improves software readability; the reduction of workforce related costs and development times, etc.

The prototype's implementation is realized in the SAP Business Warehouse environment. We chose this technology platform because SAP BW offers a comprehensive foundation for Business Intelligence tools through its architectural components: it supports the data acquisition and staging process designed to ensure quality and integration of enterprise-wide data; it enables the definition of a data warehouse layer, which stores granular, integrated data resulted from the staging processes; and it supports multidimensional data marts through an extended star schema design. We designed the prototype's main components in order to take advantage of the SAP BW capabilities and developed them for each architectural storage layer (i.e. *Data Acquisition* (data staging), *Primary Data Management* (data warehouse), and *Data Delivery* (data marts)).

Our proposed prototype enables the automated implementation of initial data warehouse and data mart structures, as well as corresponding ETL processes, from technical metadata. With its components defined for each architectural storage layer, the implementation process covers:

- data staging processes, such as replication of DataSources, generation and execution of extraction processes, generation of information modelling units (i.e. InfoObjects);
- data warehouse structures generation (i.e. DataStore objects for granular permanent data storage);
- data mart schema generation (i.e. InfoCubes for dimensional view of data);
- generation of transformations and mapping rules between the various structures and metadata objects;

The resulted initial schemas may be further enhanced through the SAP BW user interface functionality (these enhancements may include data structures re-modelling, selective data extraction, specific business-logic driven transformations, creation of routines for data cleansing and integration, etc.). We have proven that our automation prototype is beneficial in terms of cost reductions for comprehensive data warehousing solutions development, in which a large number of data structures and ETL processes are implemented through repetitive and time-consuming tasks.

Chapter 6. Conclusion and Future Work

Our main research objectives, addressed in the scope of this thesis, have been determined by the development of comprehensive data warehousing solution, driven by the cost reduction and the implementation efficiency requirements. The accomplishment of these objectives demanded a thorough understanding of the various data warehousing related aspects, namely the positioning and role of the data warehousing technology in the overall Business Intelligence framework; the definition of specific data models, which determine the storage structures of the analytical environment; the solution development methodology, guiding the effective design and implementation processes; the architecture that defines the foundation of data warehousing solution development; the framework providing a set of guidelines for building components and describing their interaction on the analytical systems; and the actual implementation processes.

The abundant amount of information existent in today's economic environment may be leveraged efficiently by means of specific applications and tools, comprising the decision support and Business Intelligence technologies. These technologies are essential in facilitating the access to enterprise-wide and external data, and providing advanced processing and analysis capabilities. As introduction into the background of the analytical systems, we began with a presentation of their main characteristics, analyzed the historical evolution of decision support applications and discussed the similarities and differences between early stage systems and nowadays comprehensive technologies.

We presented the data warehousing as a comprehensive technology used for handling the analytical environment of an enterprise, and the data warehouse as a component of the repository element and the foundation upon which Business Intelligence is built. Thus, we argued the existence of a clear differentiation between data warehousing and the data warehouse concepts, and also examined the various perspectives treated in the literature. While data warehousing comprises a series of components and processes designed to enable the collection and integration of enterprise data from various sources, with the main goal of transforming it into strategic information, the data warehouse defines the storage component of the overall technology, the repository of integrated, subject-oriented, non-volatile and time-variant data. These collected and integrated large amounts of data are subject to various mathematical and statistical calculations meant to produce valuable business insight. For this reason, the data warehouse relies on storage structures optimized for high performance querying and analysis processes, designed with specific data modelling techniques. We introduced the particularities of the various data warehousing structures (e.g. the data warehouse, the data marts), and examined the differences between them. These structures and the data of the warehouse repository are described from different perspectives by means of metadata. In order to understand their fundamental role, we presented a comprehensive study of metadata definitions, classifications and management characteristics. We focused mostly on the business and technical metadata types, highlighting the descriptive character of business metadata, essential in understanding the semantics of the business processes, and the

importance of technical metadata in enabling automation in the data warehousing environment.

Considering the complexity of the data warehousing solutions, we introduced several development methodologies aimed to offer a structured and planned approach to the numerous processes involved, by providing a repeatable, consistent and reliable set of steps or guidelines. From the various methodologies analyzed, both general and data warehousing specific, we described comprehensively two reference approaches, namely the Inmon (i.e. top-down) and the Kimball (i.e. bottom-up) models. We also introduced two frameworks defined for selecting the suitable methodology for successful data warehousing implementation, and outlined our reasons for adhering to Inmon's data-driven spiral approach in the development of our solution proposal.

The optimized data structures of the analytical environment are designed with specific data modelling techniques, meant to determine and analyze the requirements expressed by business users and to produce data models capable of supporting the enterprises' business processes. For instance, considering our comprehensive data warehousing solution approach, defined by a layer of consolidated and integrated granular data and a layer of dimensional structures, built for querying and analysis performance, we described two of the most commonly used data modelling techniques, namely Entity-Relationship and dimensional modelling. The Entity-Relationship technique presents characteristics adequate for modelling normalized models with minimum redundancy, dependencies, and inconsistencies, and thus high quality data appropriate for the data warehouse granular layer, while dimensional modelling, specific to the analytical environment, produces intuitive presentations of data in the form of querying and analysis optimized models.

Moreover, dimensional models are defined as restricted forms of Entity-Relationship models, obtained through various un-standardized methodologies (i.e. there is no universally accepted methodology as de facto standard in dimensional modelling). Considering this context, we presented a literature review of some of the most cited works in the domain, discussed their main advantages and disadvantages, and exposed our opinions regarding their appropriateness for data warehousing solution development. We selected the methodological approach proposed by Moody and Kortink [121] for our dimensional model derivation and justified our decision as follows: 1) our goal of designing a comprehensive data warehousing solution, which includes an integration layer (i.e. the central data warehouse) and a presentation layer (i.e. data marts) was fully supported by this methodology; 2) the approach is based on the enterprise data model in which data relationships are already documented, thus simplifying the data retrieval, transformation, and loading processes; and 3) the methodology has been validated in practice and it also allows the data designer to refine the development steps based on user requirements or its own business knowledge. We exemplified the methodology through a case study of a reinsurance business process modelling, derived from an Entity-Relationship schema, and contributed with enhancements regarding the business aspect of the modelling process, namely reinsurance specific modelling options, such as the analysis of the data model's utility and integration of appropriate dimensions in the dimensional model, representation of financial data through

various currencies, handling of changes in the dimensions, etc. We also evaluated the resulted data model against a series of features which models should have in order to support advanced uses in data analysis, and concluded that it complies with most requirements and thus determines a valid representation of data.

Another important aspect covered in our thesis was the selection of the appropriate data warehouse architecture. Due to the complexity and the broad scope of the data warehousing solution, its implementation architecture is significantly different and more complex than the classical database architecture. We introduced several types of architectures acknowledged in the literature (e.g. the independent data marts, the data mart bus architecture, the enterprise data warehouse, the centralized and the federated architectures) and presented two reference architectural implementations: the top-down (achieved by the Inmon model), also known as the Corporate Information Factory framework, and the bottom-up (achieved by the Kimball model), known as the Data Mart Bus Architecture. We also presented a framework defined for facilitating the selection of the appropriate architecture type, determined by organizational, environmental, project-related, technical and educational factors, etc. and selected as most suitable architecture for our solution development, the comprehensive enterprise data warehousing architecture recommended by the Inmon model.

Regarding the development framework proposed, we began from the idea that successful data warehousing design and implementation processes are supported by compatible architectures and methodological frameworks. Considering the numerous structures and processes which comprise the data warehousing solution, we defined an automated implementation framework and a corresponding prototype, based on the assumption that repetitive and time-consuming development activities may be carried out more efficiently and in a shorter period of time. We justified our proposal by presenting the various benefits of automation in software development in general, and particularly in the data warehousing environment. Our proposed framework consists of five main components, defined for data and metadata management, data staging, data warehouse and data mart schemas generation processes. We described carefully the role and specifics of each component, as well as the interaction between them. We designed the corresponding prototype for automating the creation of initial data warehousing storage structures for the enterprise data warehouse and the data mart layers, along with complementary ETL processes, from technical metadata, and accomplished its implementation in the SAP Business Warehouse environment. Our choice of technology platform was determined by the fact that SAP BW offers a comprehensive foundation for Business Intelligence through its architectural components: it supports the data acquisition and staging process, designed to ensure quality and integration of enterprise-wide data; it enables the definition of a data warehouse layer of granular, integrated data resulted from the staging processes; and it supports the definition of multidimensional data marts through an extended star schema design. Thus, we built the prototype's components in order to take advantage of the SAP BW capabilities and developed them for each architectural data storage layer (i.e. the data staging component for the *Data Acquisition* layer, the data warehouse component for the *Primary Data Management* layer, and the data marts component for the *Data Delivery* layer).

We achieved the automated implementation of an initial data warehousing schema through our prototype's deployment, as follows: the normalized flat data structures of the data warehouse, obtained as one-to-one mappings of the data sources developed on the source systems; and the relational star schema model of the data mart layer, obtained through the generation of specific data structures for our derived reinsurance dimensional data model. Nevertheless, we defined the boundaries of the automation obtained through our prototype and highlighted the fact that the resulted initial schemas may be further enhanced through the SAP BW user interface functionality (these enhancements may include data structures re-modelling, selective data extraction, specific business-logic driven transformations, creation of routines for data cleansing and integration, etc.). We also proved that our automation prototype is beneficial in terms of cost reductions for comprehensive data warehousing solutions development, in which a large number of data structures and ETL processes are implemented through repetitive and time-consuming tasks.

The results presented in our thesis have been disseminated through several articles presented at national and international conferences, and published in conference proceedings and journals of different categories. We validated our automation prototype through an article presented at the 2nd World Conference on Innovation and Computer Sciences and included in the Procedia Technology Journal, Elsevier Publishing Ltd., indexed on the ScienceDirect, Scopus and Thomson Reuters Conference Proceedings Citation Index (Web of Science) [132]. We also validated our framework proposed for the enterprise data warehouse implementation through a BDI indexed article published in the Database Systems Journal, ASE Bucharest Publishing [131]. Moreover, we discussed fundamental topics presented in our thesis, such as the importance of automation in the data warehousing environment [133], aspects of metadata modelling within data warehouses [137], comparison of methodologies used for building the data warehouse structures [135], security issues in the SAP BW environment [134] [136], etc. in several other articles, as introduced below:

- **I. M. Nagy**, *Automation prototype for the development of data warehousing data structures*, accepted for publishing in Procedia Technology Journal, Elsevier Publishing Ltd., ISSN: 2212-0173 (indexed ISI)
- **I. M. Nagy** and E. Tolea, *A Metamodel for Manipulating Business Knowledge Within a Data Warehouse*, Proceedings of the 6th International Conference On Virtual Learning, Editura Universitatii din Bucuresti, ISSN: 1844-8933, pp. 255-261 (indexed ISI Proceedings)
- **I. M. Nagy**, *A Framework for Semi-Automated Implementation of Multidimensional Data Models*, Database Systems Journal, Volume 3, Issue 2, Editura ASE Bucuresti, 2012, ISSN: 2069-3230 (indexed BDI)
- **I. M. Nagy** and C. Stefanache, *Ensuring Data Protection in the SAP Business Information Warehouse: A Case Study*, Journal of Applied Computer Science & Mathematics, Volume 9, Issue 4, 2010, ISSN:1843-1046, pp. 83 – 87 (indexed BDI)
- **I. M. Nagy** and L. Feischmidt, *Mobilizing Business Processes Security issues and advantages of using SAP Mobile Infrastructure in the development of mobile*

application, Economy Informatics, Volume 10, Issue 1, 2010, pp. 44 – 52. (indexed BDI)

- **I. M. Nagy**, *The Importance of Automation in the Data Warehousing Environment – A Case Study*, 19th International Economic Conference "The Persistence of the Global Economic Crisis: Causes, Implications, Solutions", Sibiu, 2012, pg. 201 - 208, ISBN 978-606-12-0323-9
- **I. M. Nagy** and A. Onaciu, *Two Methodologies for Deriving the Data Warehouse Structure*, Proceedings of the 2nd Symposium on Business Informatics, Austrian Computer Society Conference, pp. 198 –206, ISBN: 978-3-85403-280-9

We also contributed to a monograph on intelligent Decision Support Systems with a sub-chapter treating theoretical Business Intelligence and data warehousing aspects:

- Nițchi Ioan Ștefan, Airinei Dinu, Arba(Cordis-Herbil) Raluca, Benţa Dan, Brândaș Claudiu, Buchmann Robert, Crișan Emil Lucian, Homocean Daniel, Jecan Sergiu, Kleinhempel Simona, Mihăilă Adrian-Alin, Muntean Mihaela, **Nagy Ilona Mariana**, Petrușel Răzvan, Podean Ioan Marius, Rusu Maria Lucia, Sitar-Tăut Dan Andrei, book, *Sisteme inteligente de asistare a deciziilor*, Risoprint, Cluj-Napoca, 2010.

As directions for future work, we consider that the prototype presented as part of our contribution to the thesis may be extended to cover the automation of other structures and processes of the data warehousing environment. It may also be used per sub-modules for dividing the execution of several processes, which are, in the current context, executed for the entire data model. For instance, sub-parts of the prototype's modules may be developed and used exclusively for generating the data modelling metadata objects (i.e. InfoObjects, DataStore objects, etc.) from data architect-defined technical documents, which would greatly shorten the implementation time and thus considerably reduce related costs. Moreover, the procedures implemented for automating the technical validation of the generated data structures may be used separately for performing these actions for already existent objects throughout the entire data warehousing environment.

Additional feasible future development directions include: the partial or complete inclusion of business metadata in the automation process (e.g. in data transformation between the *Primary Data Management* and the *Data Delivery* layers); the automated generation of master data structures as part of the *Primary Data Management* layer, etc. Thus, we consider having proposed a coherent automation framework and a flexible prototype, which may bring important benefits to enterprises undertaking data warehousing development, as well as maintenance processes.