INTERNAL CONTROL RISKS WITHIN THE DATA WAREHOUSE ENVIRONMENT

by

SEAN PAUL DE LA ROSA

Submitted in fulfillment of the requirements for the degree

MAGISTER COMMERCII
(COMPUTER AUDITING)

in the

Faculty of Economic and Management Sciences

at the

UNIVERSITY OF PRETORIA

PRETORIA NOVEMBER 1999

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my study leader, Prof. Dr. H. de Jager for his ongoing support and assistance during this project.

The research is dedicated to my parents, Steve and Peggy de la Rosa for affirming my ambitions and their unending words of encouragement. All the honour and glory for both the preparation and successful completion of this work goes to my Lord and Saviour Jesus Christ who gave me the wisdom, understanding and endurance.

Thanks Tommy.

JOHANNESBURG

NOVEMBER 1999

"He that will not apply new remedies must expect new evils, for time is the greatest innovator."

Francis Bacon 1561 - 1626

"The significant challenges we face today cannot be resolved by the same level of thinking that created them."

Albert Einstein 1879 - 1955

TABLE OF CONTENTS

Chapter 4: Dependant Data Mart Environment (continued)

Chapter 5: Distributed Data Warehouse Environment 63

Chapter 6: Future Developments and Trends 76

SUMMARY

INTERNAL CONTROL RISKS WITHIN THE DATA WAREHOUSE ENVIRONMENT

by

SEAN PAUL DE LA ROSA

LEADER	:	**PROF. DR. H DE JAGER**
FACULTY	:	**ECONOMIC AND MANAGEMENT SCIENCES**
DEPARTMENT	:	**SCHOOL OF ACCOUNTANCY**
DEGREE	:	**MCOM (COMPUTER AUDITING)**

Decision support systems enable management to improve both their strategic and short term decision making processes. In most cases, decision support systems rely on historical transactional data to provide trends and statistics which management teams can base their decision making process on. An example of a decision support system is the data warehouse.

The data warehouse environment has received increased attention over the years. The augmented importance of this evolving technology has emphasised the need for management to be assured that the environment is well controlled. Management also require reliable and accurate information to be provided timeously.

The internal audit profession strives to provide management with suitable suggestions as to how internal control risks can be identified and managed within the various high risk areas of their organisations. A study of the sufficiency of audit resources relating to the data warehouse was conducted. It is evident that the internal audit profession has been slow in identifying the specific internal control risks relating to such an

environment. With this in mind, we undertook a study to determine the nature of the specific internal control risks within the data warehouse environment.

The overall purpose of the study was to provide the internal auditor with an understanding of the internal control risks within the various components of the data warehouse. Suitable internal control considerations to assess the extent of such risks were also outlined.

The study also provides a brief insight into what future trends and developments can be expected within the data warehouse environment. The effect that such enhancements could have on the internal auditor's assessment of internal control risk were also noted.

The study concludes by providing four general recommendations which should minimise internal control risks to a satisfactory level. Provided that these recommendations are considered, both management and the internal auditor will achieve maximum benefit from this environment.

OPSOMMING

INTERNE BEHEER RISIKO'S IN DIE INLIGTINGSTOOROMGEWING

deur

SEAN PAUL DE LA ROSA

LEIER	:	**PROF. DR. H DE JAGER**
FAKULTEIT	:	**EKONOMIESE EN BESTUURSWETENSKAPPE**
DEPARTEMENT	:	**SKOOL VIR REKENMEESTERSOPLEIDING**
GRAAD	:	**MCOM (REKENAAR OUTIDEERING)**

Besluitneming-ondersteuningstelsels maak dit moontlik vir bestuur om beide hul strategiese en kort termyn besluitnemingsprosesse te verbeter. In die meeste gevalle, maak die besluitneming-ondersteuningstelsels staat op histories transaksiedata om bestuurspanne van tendense en statistieke te voorsien waarop hulle hul besluitnemingsprosesse kan baseer. 'n Voorbeeld van 'n besluitneming-ondersteuningstelsel is die inligtingstoor.

Die inligtingstooromgewing het oor die afgelope paar jaar toenemende aandag geniet. Die toenemende belangrikheid van hierdie ontwikkelende tegnologie het die behoefte beklemtoon dat bestuurslui seker moet wees dat die omgewing goed beheer word. Bestuur verlang verder om betyds van betroubare en akkurate inligting voorsien te word.

Die interne oudit professie streef om bestuurslui met gepaste metodes te voorsien waarop interne beheer risikos geïdentifiseer en bestuur kan word, met inagname van die verskillende hoë risiko-areas in hul organisasie. Gebaseer op 'n studie oor die

beskikbaarheid van oudit hulpbronne wat verband hou met die inligtingsstoor, blyk dit die professie traag is met die identifisering van spesifieke interne beheer risiko's wat verband hou met so 'n omgewing. Met hierdie waarneming in gedagte, is 'n studie onderneem met die doel om die kenmerke van die spesifieke interne beheer risiko's binne die inligtingsstoor omgewing te bepaal.

Die algehele doelwit van die studie was om die interne ouditeur se kennis en begrip uit te brei aangaande interne beheer risiko's binne die verskillende komponente van die inligtingstoor. Gepaste interne beheer-oorwegings om die erns van sulke risiko's te evalueer, is ook uitgestip.

Die studie verskaf ook 'n kort insig oor watter toekomstige tendense en verwikkelinge verwag kan word binne die inligtingsstooromgewing, asook die effek wat sulke verbeteringe kan hê op die interne ouditeur se evaluasie van interne beheer risiko's.

Die studie sluit af deur vier algemene aanbevelings te maak wat sal verseker dat interne beheer risiko's sal verminder tot 'n meer aanvaarbare vlak. Indien hierdie aanbevelings oorweeg word, sal beide bestuur en die interne ouditeur maksimum voordeel uit hierdie omgewing kan put.

LIST OF FIGURES

Chapter 1

Introduction

1. Background

In today's highly competitive business markets where survival of the fittest is considered the norm, business management seek tools and new strategies which allow them to stay one step ahead of their competitors. Decision support systems are yet just another tool identified by management teams in attaining this objective (Weber, 1982: 117). This evolving technology allows for the cultivation of information-sharing within organisations thereby enabling employees to solve dynamic problems and reduce costs (Warigon, 1998: 55).

The ability of decision support systems to be scaled down or up to meet their user's needs is another primary reason why decision support technology not only favours large conglomerates (Murphy, 1997: 1). Users processing a hundred to a thousand transactions a month derive just as much benefit from the decision support system as a multi-national processing millions of transactions a day. These systems may be heavily relied upon by management in deriving strategic and operational management decisions. Therefore both internal auditor and systems developer have significant responsibilities to ensure that management are aware of the consequences should continuity of operations or the quality of data be jeopardised (Curtis & Joshi, 1997: 40).

In a survey conducted by a South African computer magazine (Du Plessis, 1998: 1), data warehousing was identified as one of the top seventeen application and system areas which information technology specialists acknowledged were crucial to the survival of their organisations. In addition, another computer article (Anon, 1998: 22) cited the data warehouse market being worth $1.47 billion in 1996 with an estimated growth to 1998 of 28%. The article went on further to indicate that the increased demand was for applications that enabled organisations to access, analyse and report data.

Internal auditors seek to align themselves with the organisation's major objectives and focus on adding value to the business process. They therefore have a unique responsibility to focus on what management consider critical to the organisation's success and overall survival (Ridley, 1996: 24). If management considers the data warehouse environment to be a significant mainstay within their operations, internal auditors own the responsibility of ensuring that they are able to advise management concerning the internal control risks within such an environment.

2. Purpose and reason for study

2.1 Purpose of study

The purpose of this study is to identify the most pertinent internal control risks within the data warehouse environment. The study will also suggest suitable internal control considerations which the internal auditor can apply in assessing the extent of such internal control risks. Finally, the impact of future developments within the data warehouse environment on the assessment of internal control risks will be considered.

2.2 Reason for undertaking the study

Based on an extensive evaluation of audit resource materials, it would seem that little attention has been given to the data warehouse environment by the internal audit profession. There has also been a lack of focus on what impact this evolving technology will have on the assessment of internal controls[1]. The study will address internal control risks at the following stages:

- Development phase of the data warehouse.
- The established data warehouse environment.
- Dependent data mart.
- Distributed data warehouse environment.
- Future developments within the data warehouse environment.

[1] *Assumption based on the writer's personal evaluation of internal and external audit source materials and the results of extensive world wide web searches relating to the audit of the data warehouse environment.*

3. Defining the data warehouse environment and internal control risk

The following section provides an indication of how the data warehouse forms part of the decision support environment. It also describes the fundamental principles of the data warehouse. As an introduction to the remainder of the study, this section will conclude by identifying the overall internal control risks that can be expected within the data warehouse environment.

Since internal control risk is the central theme of this study, we will first give attention to defining internal control risk.

3.1 Defining internal control risk

Internal control risk is the risk that management's plans, organisation and associated procedures will not provide reasonable assurance that the organisation's goals and objectives will be achieved (IIA, 1983). We will also rely on COBIT's (ISACA, 1998) information criteria identified by the Information Systems Audit and Control Association as a method of categorising risk exposure to the organisation. These criteria are:

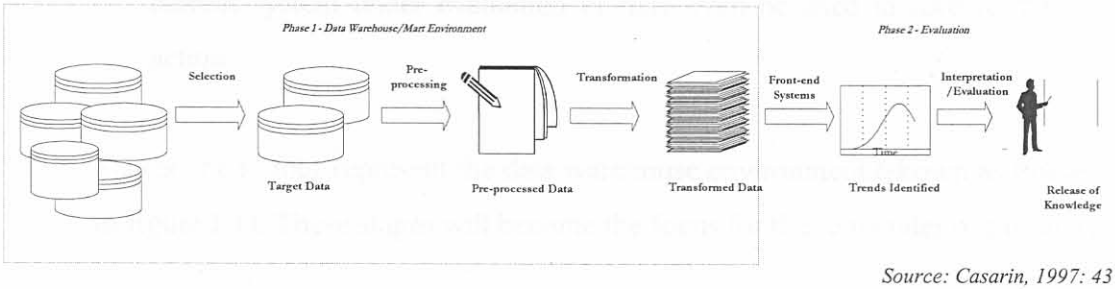
- *Effectiveness:* Information being relevant and pertinent to the business process as well as being delivered in a timely, correct, consistent and usable manner.
- *Efficiency:* Provision of information through the optimal (most productive and economical) use of resources.
- *Confidentiality:* Protection of sensitive information from unauthorised disclosure.
- *Integrity:* Accuracy and completeness of information as well as to the validity in accordance with business values and expectations.
- *Availability:* Information being available when required by the business process now and in the future. It also concerns the safeguarding of necessary resources and associated capabilities.
- *Compliance:* Complying with those laws regulations and contractual arrangements to which the business process is subject, i.e. externally imposed business criteria.

- *Reliability*: Provision of appropriate information for management to operate the entity and for management to exercise its financial and compliance reporting responsibilities.

3.2 Interaction with the decision support environment

It is important that the internal auditor understands how the decision support environment is structured so that he/she will be able to effectively identify and assess internal control risk.

Figure 1.1 - Overview of the knowledge discovery in databases process



Source: Casarin, 1997: 43

A process referred to as the Knowledge Discovery in Databases was developed to fully explain this environment. Figure 1.1 above provides a brief overview of the Knowledge Discovery in Databases process (Fayyad, 1996: 27-34).The Knowledge Discovery in Databases process is split into five individual stages (Casarin, 1997: 43-46):

- *Selection*

  This stage involves identifying data which is needed by management to aid in the decision making process. It also identifies what decisions will be made in utilising this data.

- *Pre-processing*

  Once data is identified, the cleaning of data is performed. This is done to remove irregularities or inconsistencies which may render the data unreliable for use by management. Possible irregularities could include, missing data fields, invalid characters loaded and even duplicate records.

- *Transformation*

Data is reduced to a uniform source which is consistent in all aspects and which is considered reliable in making management decisions. This stage usually involves relocating cleansed data to a separate database.

- *Information access layer systems*
  These systems extract the uniform data and present it in a format which will aid users in either confirming or disproving their hypothesis. A large number of systems exist, each providing their own unique capabilities.
- *Interpretation and evaluation*
  Interpreting presented results creates a higher level of user knowledge. Interpreted results or newly gained knowledge can be incorporated into the current system under evaluation or may even be used to take formalised action.

Stages one to four represent the data warehouse environment (shown as Phase 1 in figure 1.1). These stages will become the focus for the remainder of the study.

3.3 The data warehouse and its concepts

A data warehouse is an architecture for organising information systems. This technology has stemmed from repeated attempts by various researchers and organisations to provide businesses with flexible, effective and efficient means of obtaining sets of data. These sets have come to represent one of the organisations most critical and valuable assets (Gupta, 1997: 15). W.H. Inmon, acknowledged as the father of the data warehouse concept, defined a data warehouse as a "subject-oriented, integrated, time variant, non-volatile collection of data in support of management's decision making process" (Inmon, 1996: 33).

This collection of data could contain both highly detailed and summarised historical data relating to various processes within the organisation. Data is stored in time frames (e.g. month, years, etc.), so that trends such as market inclinations can be spotted and production accordingly amended. The following key concepts are raised in light of Inmon's definition of a data warehouse (Inmon, 1996: 33-37):

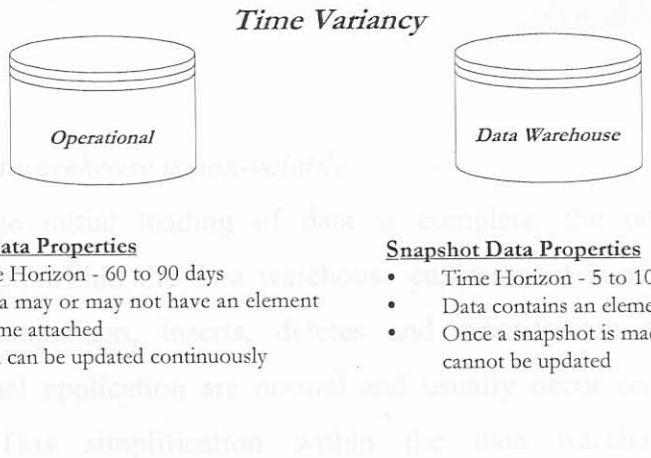
- *The data warehouse is subject orientated*

  Data stored within a data warehouse is physically separated from its source. The source which is operational in nature could include data relating to anything such as aviation records, hamburger sales or even telephone calls placed by subscribers. By nature the data warehouse supports the organisation's core functions (i.e. what the business does to derive income), and is not suited to provide information on ancillary processes within the organisation. Because of this orientation, the design and implementation of the data warehouse is affected most greatly by the major core functions and to a lesser degree by smaller, less prevalent subject areas.

- *The data warehouse has integrated components*

  All data stored must be integrated to ensure the success of the data warehouse. Naming conventions, measurements of variables and encoding structures must be consistent and no variations should be able to be entered into the data warehouse's repository. An example of the possible variations which could result in incorrect source data include the use of "M" and "F" to denote gender, while in other source applications the use of "x" or "y" may be applied. Integration affects almost every aspect of the data warehouse development. As we will see later, this issue has caused major concern not only for the internal auditor but also senior management. They run the risk of placing reliance on faulty information and trends.

Figure 1.2 - The issue of time variancy

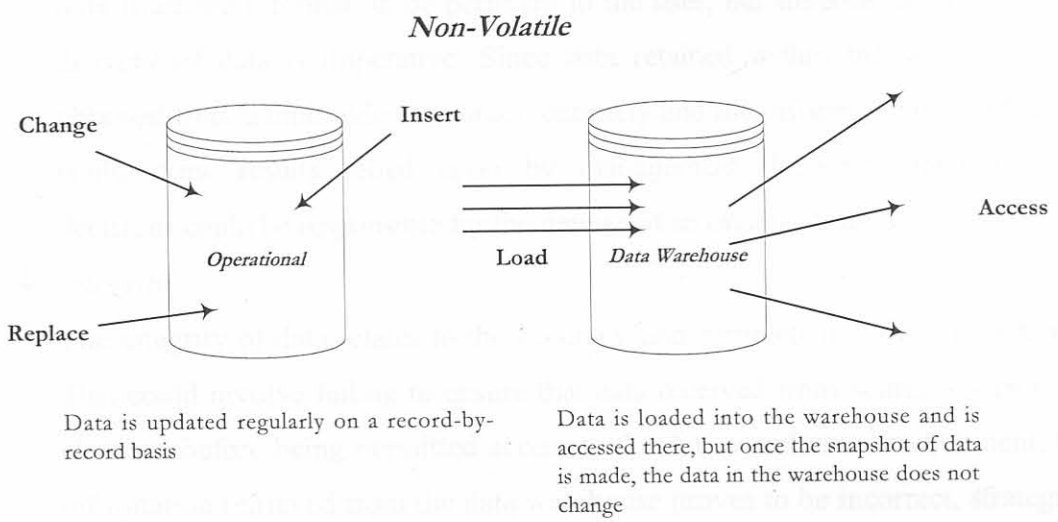


Source: Inmon, 1996: 37

- *The data warehouse is time variant*

  As detailed in Figure 1.2, the data retained within a data warehouse spans over a long time horizon - anywhere from five to ten years. Conversely in traditional applications, current values cover nothing more than sixty to ninety days (this is considered sound design for applications). Every group of data retained within the data warehouse includes some element of time, such as day, week, month, etc. This time element may even be implicit in nature. Many of the information technology professionals refer to the data warehouse as a long series of snapshots. This is based on the fact that these snapshots cannot be changed once correctly embedded within the repository (in certain instances changes to data correctly stored may be unethical).

Figure 1.3 - The issue of non-volatility



Source: Inmon, 1996: 36

- *The data warehouse is non-volatile*

  Once the initial loading of data is complete, the only other operation performed within the data warehouse environment is access to the data. In stark comparison, inserts, deletes and amendments to data within the traditional application are normal and usually occur on a record-by-record basis. This simplification within the data warehouse expedites the development process in that control over changes to data is no factor. This simplification also diminishes the need for highly complex technology

supporting backup and recovery that would usually be the case in traditional applications (figure 1.3 illustrates this).

3.4 *General internal control risks within the data warehouse environment*

According to the COBIT control criteria identified above, the major internal control risks which could be faced by an organisation employing a data warehouse, could have material effects on the successful operation of the organisation. An indication of the major internal control risks faced by an organisation embarking on an implementation or who might currently possess such an environment (Curtis & Joshi, 1997: 40-44):

- *Effectiveness*

  Effectiveness is arguably the most important criteria for a data warehouse. Not only must the information be pertinent to the user, but timeous and consistent delivery of data is imperative. Since data retained within the warehouse is obtained from a multitude of sources, untimely and inconsistent delivery of data could skew results relied upon by management. Resultant misinformed decisions could be responsible for the demise of an organisation.

- *Integrity*

  The integrity of data relates to the accuracy and completeness of information. This could involve failing to ensure that data received from source systems is cleansed before being permitted access to the data warehouse environment. If information retrieved from the data warehouse proves to be incorrect, strategic management decisions derived from such inaccurate data could also negatively impact the organisation's operations.

- *Availability*

  Since data warehouses usually retain unbelievably high volumes of data which are combined to present consolidated results, ineffective storage could result in management not being able to obtain data which can be effectively analysed. A similar scenario is true if one considers that availability is also impacted by ineffective backup and comprehensive disaster recovery procedures.

- *Efficiency*

  As management seek to minimise costs and reach optimal return on investment, warehouses can become uneconomical over time if not effectively

monitored for efficiency. Inefficient data warehouses storing needless data can adversely impact the ability of the database resources to make information available in the form and time frame needed by the user.

- *Confidentiality*

  The data warehouse is a clearly categorised repository of data, usually storing information by degree of importance. Intruders therefore find this environment irresistible. Leaking of past patents, company strategies and other sensitive information could leave an organisation doomed. Physical security risks may also negatively impact the data warehouse environment.

- *Reliability of information*

  Ineffective planning of the data warehouse could leave management with a model which does not supply the information required for improved management decision making. Not only would this result in a waste of valuable company resources, but will also deter future management from ever considering data warehousing as a tool which could improve their strategic thinking.

4. Research methodology

The research methodology applied in this study consisted mainly of an understanding of literature in the fields of internal auditing and systems and applications development.

To identify the most pressing issues within the data warehouse environment, an empirical study has also been conducted.

5. Presentation approach

The study shall comprise seven chapters which are described as follows:

- *Chapter 1: Introduction*

  The need, purpose and research methodology for the study is sketched briefly. Internal control risks, purpose, research and presentation approaches have also

been formalised. The reader is introduced to the data warehouse concept and its unique components.

- *Chapter 2: Data warehouse development*

  Internal control risks specific to the development and implementation of a data warehouse are identified. The system development methodology utilised in this development are commented on. The chapter provides the internal auditor with suitable internal control considerations which could be used to assess internal control risks. The results of the empirical study specific to data warehouse development are also presented.

- *Chapter 3: Established data warehouse environment*

  Given the nature of an established data warehouse environment, consideration will be given to highlighting the internal control risks specific to such an environment. The chapter will conclude by discussing internal control considerations which can be applied in assessing such an environment. The results of the empirical study specific to established data warehouse environment are also presented.

- *Chapter 4: Dependent data mart environment*

  The dependent data mart is a sub-environment of the data warehouse. Attention will be given to providing the internal auditor with an understanding of how the dependent data mart relies on the existing data warehouse. It also outlines what control risks exist within such an environment. The chapter will detail internal control considerations which the internal auditor can apply in ascertaining whether suitable control measures are in place to mitigate significant exposures.

- *Chapter 5: Distributed data warehouse environment*

  Internal control risks specific to the distributed data warehouse environment will be discussed. This will provide the internal auditor with in insight on how to ensure integrity of data across open communication mediums. Appropriate internal control considerations and the results of the empirical study are also provided.

- *Chapter 6: Future developments and trends*

  The chapter includes a brief insight into future developments expected within the data warehouse environment and what affect these changes could have on the internal auditor's assessment of internal control risk.

The chapter will also detail the advantages internal auditors can realise in utilising data warehouse technology as part of other audit reviews performed. The chapter will conclude by comparing how important South African internal auditors regard the data warehouse environment to other technologies in the future.

- *Chapter 7: Conclusion*
  Conclusions arising from this study are elaborated on and overall recommendations are provided. Finally, further areas of research are proposed.

6. Summary

In this chapter we identify the purpose of the study as being the identification of the most pertinent internal control risks within the data warehouse environment. This study will also provide the internal auditor with suitable internal control considerations on how to assess internal control risk.

The chapter also defines the core components of the data warehouse environment and its association with decision support systems. Finally, a brief summation of the key internal control risk areas which could exist within the data warehouse environment has been presented.

7. Conclusion

Since internal auditors have a responsibility of ensuring that management implement controls which allow them to attain their overriding goals and objectives, internal auditors should be aware of the risks and necessary control mechanisms needed to mitigate exposures within the data warehouse environment.

It is clear from statistics presented in the background section of this chapter, that the prevalence of data warehouses will dramatically increase in time to come. It is therefore imperative that the data warehouse environment be considered as an integral part of the internal auditor's universe and should be consistently monitored by management for possible control weaknesses.

Chapter 2

Data Warehouse Development

1. Introduction

Effective project management is a key ingredient for ensuring that a well controlled data warehouse is provided to the end-user. It is vital that appropriate planning and preparation take place before embarking on such an extensive exercise. If appropriately planned, the project will result in significant returns and a more controlled data warehouse environment (Kachur, 1999: 4).

2. Aim

The aim of this chapter is to identify the internal control risks which could arise during the development of the data warehouse environment. It also provides suitable internal control considerations which can be applied in assessing such internal control risks.

The development cycle for the data warehouse differs significantly from the traditional system development life cycle applied in other application and system developments. We will first consider why such differences exist (figure 2.1 reflects the traditional system development methodology which will be referred to).

The remaining portion of the chapter identifies internal control risks within the following two phases of the data warehouse development (Inmon, 1996: 73):

- Interface development between existing operational sources and the data warehouse package.
- The data warehouse package and vendor evaluation.

The chapter concludes with results of the empirical study relating to internal control risks during the development phase.

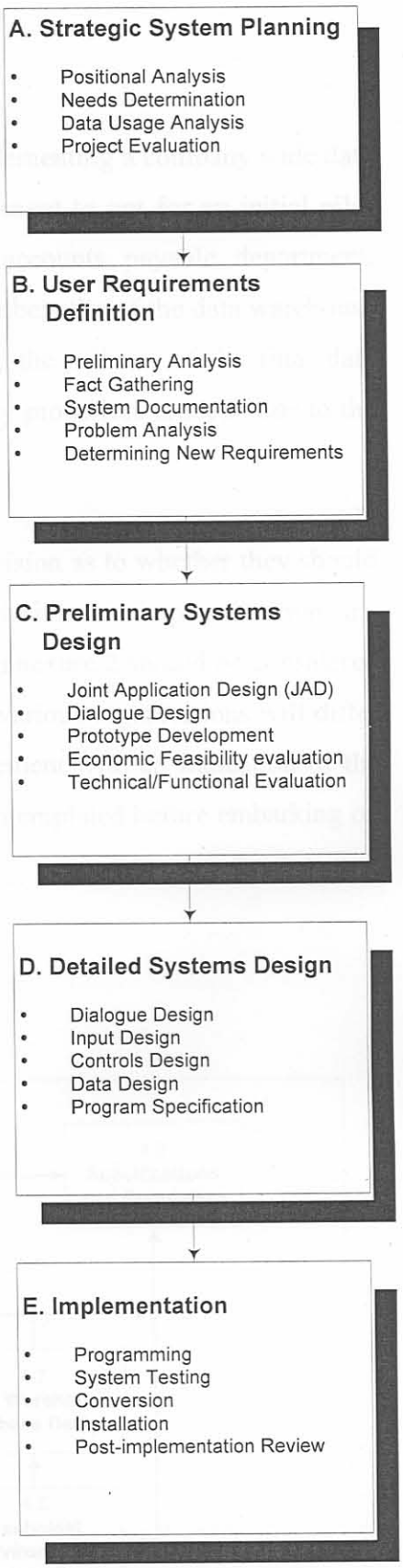
3. Why a different system development life cycle exists for data warehouses

The data warehouse development is heuristic (Inmon, 1996: 73). This means that the development and criteria of the subsequent phases of such a project are dependent on the outcome/results of previous phases within the development cycle. It is with this in mind, that the internal auditor must realise why the traditional system development life cycle cannot be applied to the data warehouse development: The exact usage requirements for the data warehouse will not be known until the data warehouse environment has been populated with data. Although management and the development team may estimate what usage they expect to derive from the data warehouse environment, they must avoid making detailed assessments until populated data has been made available (ibid.).

4. Interface development

Inmon's data warehouse development life cycle is reflected in figure 2.2. The study relies on Inmon's framework as a means of identifying internal control risks and suitable internal control considerations. The remainder of this section is structured according to the stages outlined in figure 2.2.

Figure 2.1 - Traditional system development life cycle



Source: Lay, 1993: 191

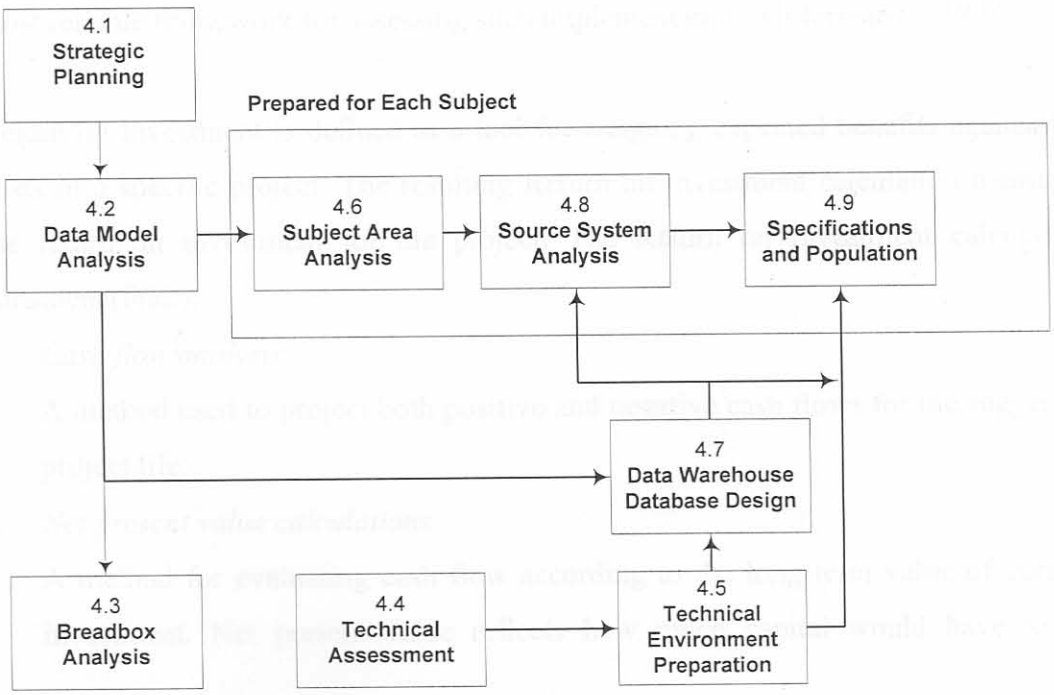
4.1 Strategic planning

4.1.1 Process steps

In instances where the organisation is considering implementing a company wide data warehouse, it is considered good practice for management to opt for an initial pilot project addressing a single operational unit, e.g. accounts payable department, strategic operations, etc. (Inmon, 1996: 298). Once the benefits of the data warehouse have been realised in the single organisational unit, the roll-out of the final data warehouse elements to the rest of the organisation may prove less arduous due to the increased user buy-in.

To assist management in making a more informed decision as to whether they should opt for a pilot project or an organisation-wide data warehouse implementation, the completion of a checklist similar to that reflected in Annexure 2 should be considered (Adelman, 1998: 1-4). Although the overall needs of various organisations will differ quite significantly, the checklist will provide management with an indication of the risk, cost and time considerations which should be contemplated before embarking on

Figure 2.2 - System development life cycle for the data warehouse



Source: Inmon, 1996: 350

a first time installation.

It is vital that justification for such a project is meaningful, clear and accurate. Implementation must successfully appear to ensure that resources will be effectively utilised and that management decision making will be enhanced.

An incomplete or inaccurate justification for the data warehouse development could result in expected benefits not being fully realised by the organisation. Some of the most probable justifications may include (Greenfield, 1998: 1-2):

- To perform querying and reporting tasks on a platform separate from that of the operational system.
- To provide an environment which will improve knowledge sharing without the need for detailed technical knowledge.
- Access a vast array of data compiled from multiple sources.
- Archive data.
- Limit access to data.

In addition to a sound justification, senior management will require a formalised cost justification as basis to deciding whether to adopt such an environment or not. The traditional project administration framework, Return on Investment is considered the most reliable framework for assessing such implementations (Informatica, 1998: 2).

Return on Investment is defined as a tool for weighing expected benefits against the costs of a specific project. The resulting Return on Investment calculation measures the return on investment for the project. The Return on Investment calculation considers (ibid.):

- *Cash flow analysis*
  A method used to project both positive and negative cash flows for the suggested project life.
- *Net present value calculations*
  A method for evaluating cash flow according to the long term value of current investment. Net present value reflects how much capital would have to be

invested currently, at an assumed interest rate, in order to create a stream of payments over time.

- *Return on investment*
  This calculation identifies the net present value of total incremental cost savings and revenue divided by the net present value of total costs multiplied by 100.
- *Payback calculations*
  A calculation showing how much time will pass before an initial capital investment is recovered.

The 1996 IDC report titled, *The Foundations of Wisdom - A Study of the Financial Impact of Data Warehousing* (Informatica, 1998: 7) indicated that among the fifty companies surveyed, an average three year Return on Investment of 401 percent was reached. Average payback for data warehouse applications was 2.3 years at an average data warehouse cost of $2.2 million.

Return on Investment frameworks do however have two significant weaknesses:

- The framework can only predict measurements for those benefits that are tangible, such as money saved, hours reduced or reports generated, etc.
- The framework cannot convey the value of what might be considered more strategic benefits, such as faster access to customer information, or making better informed business decisions.

The power in overcoming the limitations of Return on Investment frameworks lies in the ability of senior management to fine tune models regularly by replacing assumptions with actual statistics (Informatica, 1998: 7).

After the approval of the data warehouse project, management should appoint a project team as part of the strategic planning phase. The project team should consist of a project leader, business analysts, data administrators, database administrators, systems support, computer programmers and end users. These personnel will be responsible for the overall project administration, design of the warehouse structures; analysis of source data; identification of how data is to be linked and, if applicable, integration external sources (Inmon, 1996: 295).

It is very important to distinguish data administration from database administration. Data administrators are business oriented, focused on the meaning and use of data. Database administrators are however technically oriented, and are concerned with the reliability, integrity and performance of database applications. While a database administrator typically corrects application errors due to database processing problems, a data administrator deals with business problems due to incorrect data values or invalid use of data (Lambert, 1998: 7-9). A detailed breakdown of the roles and responsibilities of a data administrator has been included under Annexure 3.

4.1.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, risks within the above mentioned process affects the reliability criteria of information.

The following detailed internal control risks are identified:

- By not adopting a project framework specific to the data warehouse development, the efficient and effective implementation of the data warehouse will be hampered.
- Resources will be wasted if an initial pilot project is not run to ascertain whether data warehouse benefits will be realised on a smaller scale.
- Incomplete or inaccurate project justifications could result in the organisation not realising the expected benefits of the data warehouse development.
- An inaccurate cost administration framework could result in expected project benefits not being clearly identified and accounted for.
- The appointment of an incomplete project team could result in key processes during the data warehouse development not being addressed (this includes the establishment of the data administrator role within the organisation).

4.1.3 Internal control considerations

The following internal control considerations are applicable:

- An accepted data warehouse development framework should be adopted by the project team.

- An initial pilot project to ascertain whether data warehouse benefits will be realised on a smaller scale should be considered before an organisation wide data warehouse is implemented.
- A complete and accurate project justification should be prepared. It should identify the exact reasons for the suggested implementation of the data warehouse.
- By justifying the project, the team can analyse and document the security threats, potential vulnerabilities and their impact.
- A comprehensive cost administration framework, such as Return on Investment, should be implemented.
- The cost administration framework should provide for an analysis of the costs and benefits associated with each alternative being considered for satisfying the established business requirements.
- A complete project team should be appointed.
- The data administration function should be established and the roles and responsibilities of the function clearly defined.

4.2 Data model analysis

4.2.1 Process steps

Inmon (Inmon, 1996: 81) indicates that two generic methodologies exist which are applied in the development of applications. These models are the data and process model. It is important to distinguish between the two models since the application of the incorrect type could prove costly in terms of human resources and time delays.

Inmon is of opinion that the process model is not effective in the development of a data warehouse. This is because most traditional applications have specific deliverables and functions which must be provided to the user. The data warehouse is however a neatly compiled source of data which can be utilised in a diverse number of management hypothesis. It is even possible that by applying the process model, that the Information Technology department will limit the true functionality which could be provided by a data warehouse. Accordingly, Inmon suggests that the

Information Technology department focus on the data model as a suitable approach in the development of the data warehouse (Inmon, 1996: 73-74).

The ultimate aim of the data model is to ensure that the major subject areas have been identified. The data model is split into three distinct levels, viz. high, middle and low level models (Inmon, 1996: 85 - 96):

- The high level model defines the boundaries in which the data warehouse will operate, i.e. which application's data will be included in the data warehouse and which data classes will be left out. The model includes details on the keys and types of data classification. A key is a data element or combination of data elements used to identify or locate a record instance. A key may be primary or secondary.
- A middle level model is developed for each application or subject area defined in the high level model. This process usually begins by separating primitive and secondary data. Primitive data is defined as data elements whose existence depends on a single occurrence of a major subject area of the enterprise. Secondary data is defined as data elements whose existence depends on two or more occurrences of a major subject. This distinction is made to ensure that duplicate data elements are avoided and that the most accurate data element is selected if duplicates are found. This model is concluded once the project team have identified the relationships between the primitive and secondary data classes.
- The low level model is obtained by expanding the middle level model to include detailed information relating to each key and physical hardware considerations. Hardware considerations are affected by the level of detail contained in the unit of data (often termed granularity) as well as the technique used to divide data into physical units so as to improve overall performance.

4.2.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, the risk within the above mentioned process affects the integrity and reliability criteria of information.

The following detailed internal control risk is identified:

- The data model must be applied in the development of the data warehouse. If not, the project team may not detect all major subject areas and ensure that the data warehouse relies on the most accurate data available from various source systems.

4.2.3 Internal control considerations

The following internal control considerations are applicable:

- The project team should adopt the data model as a preferred framework for the development of the data warehouse.
- The data model should be defined in such a way that it includes:
  i. High, middle and low level sub-models.
  ii. Identifies major subject areas.
  iii. Clearly defines the boundaries of the model.
  iv. Separates primitive and secondary data.
  v. Keys, attributes, data relationships and duplicate data for each subject area are identified.

4.3 Breadbox analysis

4.3.1 Process steps

After the data model has been finalised, the project team will need to determine the volume of data which will be retained within the data warehouse environment. The Breadbox Analysis simply projects a rough estimate of how much data the data warehouse will hold (Inmon, 1996: 336).

The Breadbox Analysis is initiated by estimating the number of rows of data which will be housed in the data warehouse. Inmon (Inmon, 1996: 145) defines an algorithmic path which should be used in calculating the space needed to retain the data records. The calculation involves determining the number of expected rows for each known table as well as what the maximum and minimum number of rows in the data warehouse. The calculation is concluded by multiplying the biggest and smallest

row estimates (in terms of bytes) by the number of individual maximum and minimum rows for the first year.

From this, it is apparent that the necessary space needed to house the data is directly affected by the granularity of the data. The more detailed data which must be retained reflects a lower level of granularity as opposed to a high level of granularity which indicates more summarised data (ibid.).

In instances where the data warehouse needs to contain a large volume of data, multiple levels of granularity will need to be considered (Inmon, 1996: 336). If the data warehouse is not going to contain a massive amount of data, then there is no need to plan a design for multiple levels of granularity.

4.3.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, the risk within the above mentioned process affects the integrity and reliability criteria of information.

The following detailed internal control risk is identified:

- An incomplete assessment of the volume of data which will be retained by the data warehouse could result in:
  i. Insufficient or excessive hardware being purchased.
  ii. User needs not being met (by not ensuring that a sufficient level of granularity has been taken into account).

4.3.3 Internal control considerations

The following internal control considerations are applicable:

- The project team should have completed a formalised Breadbox Analysis before preceding further with the data warehouse development.
- The project team has to have defined the total processing requirements for the data warehouse environment at maturity.

- The method adopted by the project team in determining the necessary hardware and software capacity should be based on the Breadbox Analysis.

4.4 Technical assessment

4.4.1 Process steps

This phase focuses on determining the architectural configuration needed for the data warehouse. No pre-defined format for the assessment is proposed, since it will depend on whether the organisation decides to house the data warehouse on existing hardware or on newly purchased equipment. If executed properly, the technical assessment will address the following criteria (Inmon, 1996: 337):

- The ability to manage large amounts of data.
- The ability to allow data to be accessed flexibly.
- The ability to receive and send data to a wide variety of different platforms for further use.

4.4.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, the risks within the above mentioned process affect the availability and reliability criteria of information.

The following detailed internal control risks are identified:

- An incorrect technical assessment could result in the final data warehouse not being able to handle the expected data volumes.
- Information will not be delivered to the user on a timely and consistent basis.

4.4.3 Internal control considerations

The following internal control considerations are applicable:

- A technical assessment addressing data volumes should be prepared.

- Management should have implemented suitable scaling procedures to manage the expected increase and shrinkage in data volumes over time (Pine Cone Systems, 1996: 6).
- The final assessment should be compared against industry standards as a means of benchmarking the assessment's appropriateness and accuracy.
- In instances where the organisation has opted for newly purchased equipment, the project team should take steps to ensure that lead times for providing the equipment are in line with the suggested project plan.

4.5 Technical environment preparation

4.5.1 Process steps

Once a suitable architecture has been defined, the project team will need to identify how it will be accommodated. This technical phase will ensure the following issues are addressed (Inmon, 1996: 338):

- How the organisation's Information Technology network will be affected by the increased traffic due to the data warehouse environment.
- The nature of traffic, either short or long bursts, generated by the data warehouse.
- How to minimise and/or alleviate processing conflicts between the organisation's existing applications and the data warehouse.

4.5.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, the risk within the above mentioned process affects the effectiveness and availability criteria of information.

The following detailed internal control risk is identified:

- The organisation's existing applications and Information Technology operations could be negatively impacted by the introduction of the data warehouse environment if the technical environment is not suitably prepared for the architectural configuration.

4.5.3 Internal control considerations

The following internal control considerations are applicable:

- The project should confirm that the expected increase in network traffic will not affect the operation of other critical applications currently in use.
- Suitable monitoring procedures should be implemented by the project team thereby ensuring that increases in network traffic are identified timeously and corrective procedures initiated.

4.6 Subject area analysis

4.6.1 Process steps

This is defined as the first experimental phase of the development process. The subject area analysis follows on from the data model analysis and will identify suitable population data from existing applications which should be introduced into the data warehouse environment (Inmon, 1996: 280-281). Subject areas could include: customer details, product archives, account histories, transaction activity records, shipment trails, etc.

The success of the subject area analysis is directly affected by the accuracy and comprehensiveness of the data model analysis. It is considered good practice to introduce a subject area large enough to be meaningful and small enough to be implemented (Inmon, 1996: 339). the project team must start with the completed data model and asks what data is in hand that best fulfills the data requirements identified in the data model (Inmon, 1996: 278). This is to ensure that the data warehouse environment provides the most reliable information to the end user.

The project team may encounter instances where the exact type of data specified in the data model cannot be located within any one specific subject area. In such instances, the team will either need to develop suitable profile records which

aggregate loose records from various sources (Inmon, 1996: 122-124), or select another subject area for development.

4.6.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, the risks within the above mentioned process affect the effectiveness and reliability criteria of information.

The following detailed internal control risk are identified:

- The user's needs will not be met if the most appropriate subject areas are not chosen, based on the details contained in the data model analysis.
- Inaccurate and untimely data reliance could occur if the most appropriate subject data is not selected in instances where subject data can be retrieved from multiple sources.

4.6.3 Internal control considerations

The following internal control considerations are applicable:

- The chosen subject areas should agree to those previously defined in the data model analysis.
- The project team should take steps to ensure that the initial subject area selected is small and meaningful enough to ensure implementation success.
- In instances where subject data can be retrieved from multiple areas, the project team take steps to ensure that the most timely, complete and accurate source system is chosen.

4.7 Data warehouse design

4.7.1 Process steps

The completed data model and subject area analysis are critical to the success of an accurate data warehouse design (Inmon, 1996: 278). To provide the project team with

a completed data warehouse design, they will need to adjust the data model analysis with the following (Inmon, 1996: 278-280):

- Data used purely for operational purposes should be removed from the analysis (this could include any form of data which will be of no benefit to the end user as part of the data warehouse environment).
- Relationships between operational data elements which ensure referential integrity, i.e. that both data elements are kept up to date with all changes made, must be removed and details kept unchanged (this is based on the premise that once data has entered the warehouse, the data does not change).
- Data which is derived from calculations and other real-time sources must be included into the design where needed.
- Data should be grouped according to their propensity for change. Often termed a stability analysis, it will aim at grouping data elements together which will change based on similar conditions.
- The data warehouse should be organised according to the subject areas initially defined in the subject area analysis.

4.7.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, the risks within the above mentioned process affects the effectiveness, integrity and reliability criteria of information.

The following detailed internal control risks are identified:

- An incomplete data warehouse design could produce incomplete and unreliable data elements in the final data warehouse environment.
- Misinformed end user decisions could result which, depending on the significance of the data warehouse query, could directly affect the viability of the organisation.

4.7.3 Internal control considerations

The following internal control considerations are applicable:

- The project team should rely upon the data model analysis as a basis for preparing the data warehouse design.
- The project team should take steps to ensure that purely operational data elements (i.e. data elements which will not benefit the end user in the data warehouse environment) are removed from the design.
- The project team should take steps to ensure that a comprehensive stability analysis is completed.

4.8 Source systems analysis

4.8.1 Process steps

Based upon the results of the subject area analysis, the project team is required to perform a source systems analysis as a basis for assessing whether the developed data warehouse environment is closely aligned with the operating systems (Kimball, 1996: 1). This analysis also assists the project team to understand how the operational systems function and how data can be effectively converted to be of maximum benefit to the user. The analysis consists of three key stages (Kimball, 1996: 2):

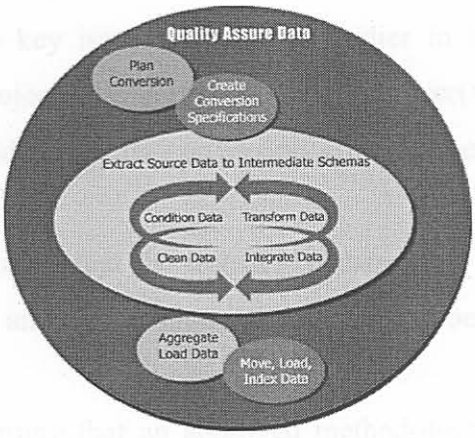
- *Definition of source data elements*
  This step attempts to verify that data labeled within the operational system retains as much of its original character as possible as origin must be easily traced once included into the data warehouse. This improves traceability and follow-up of inconsistencies in data should inaccurate data be detected.
- *Evaluating the accuracy of data before migration to the data warehouse*
  This process attempts to highlight data which is not frequently relied upon within the operational systems, but which may be used or summarised in the data warehouse. These reviews may be performed electronically or manually.
- *Managing the volume of data elements*
  Unmanaged data transfer and unnecessary data elements included in the data warehouse can result in a totally ineffective data warehouse environment. The source system analysis will ensure that only needed data elements are included in the final data warehouse.

The project team will need to prepare a detailed conversion plan in order to attain the goal of mapping the data from the operational environment to that of the data warehouse (Bohn, 1997: 1). All team participants need to understand the conversion requirements and what standards have been adopted by the organisation in the data warehouse development. This plan also identifies the best route to migrate source data to the data warehouse. The core elements of the plan should include (Bohn, 1997: 2-3):

- An identification of each source system's operating platform.
- Programmatic language of the source system.
- Access method to obtain the data from the source system.
- What considerations have been made to ensure that sufficient machine resources are in place.
- Detailed procedures on how the operational data will be transferred to an intermediate schema before transfer to the final data warehouse.
- What procedures the task team will comply with in the conditioning and transforming data, e.g. conversion specifications, the rejection of duplicate and invalid data, etc.
- What tolerance levels for incorrect data values will be accepted within the data warehouse environment.
- How the team will load and index the data in the data warehouse, i.e. a uniform naming convention is applied for data elements.
- What procedures will be applied in migrating data over to the final data warehouse.
- Data cleaning requirements.
- Detailed procedures on how end-user reviews and sign-off will be performed
- Data validation and correction procedures and the process to be applied in reconciling data to source.
- Procedures to ensure that all data transferred from the operational system is transferred in the most appropriate time frame. The project team, in conjunction with the end user, must identify the time when the upload of data should take place so as to reflect the correct data characteristics. (Inmon, 1996: 192-195).

Figure 2.3 provides an outline of a controlled data conversion process.

Figure 2.3 Controlled data conversion process



Source: Bohn, 1997: 5

We define three different types of data transfers which must be catered for as part of the interface between external systems and the data warehouse environment (Inmon, 1996: 76-80):

- Archival data loading.
- Data transferred from operational data sources (this includes both internal and external sources).
- Suitable checks which ensure that any changes made to operational data are effectively followed-up and that data within the data warehouse is corrected.

4.8.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, the risks within the above mentioned process affects the integrity, effectiveness and reliability criteria of information.

The following detailed internal control risks are identified:

- An incomplete source system analysis could result in incorrect data elements being transferred to the data warehouse.
- Non-existence of a conversion plan could result in team participants being unaware of the approved standards and conversion requirements which should be applied in the development of the data warehouse.
- Without a source systems analysis, it may prove difficult for the project team to trace incorrect data residing within the data warehouse environment back to the operational system that created the errant data.

4.8.3 Internal control considerations

The following internal control considerations are applicable:

- The project team should have completed a detailed source systems analysis, including a comprehensive conversion plan.
- The conversion plan should address the key issues highlighted earlier in this section and should be signed off by the project team participants and end user(s).
- Suitable procedures should be established to handle conversion errors detected during the migration process.
- The project team should take steps to ensure that the transfer of data from the operational system will first reside in an intermediate schema before being transferred to the data warehouse.
- The project team should take steps to ensure that an approved methodology is complied with in updating data already resident within the data warehouse with changed data in the operational systems.

4.9 Interface specifications and population

4.9.1 Process steps

This stage involves the following activities (Inmon, 1996: 280):

- Actual condensation of data thereby removing all unnecessary data elements as predefined in the subject area analysis.
- Fixing the time basis on which data should be refreshed.
- Final integration of data from the systems and application-orientated environments.
- Executing the technical procedures as reflected in the approved conversion plan.
- Establishing conversion specifications.

Subsequent to the completion of the conversion plan, the project team will need to develop the conversion specifications (Bohn, 199: 3). The conversion specifications reflect source data maps linking data elements from the operational systems to the data warehouse.

After sign-off of the conversion specifications, the project team will need to develop the programmatic code needed to transfer data from the various source systems to the

data warehouse environment. Programmatic coding consists of six types of routines that perform the extraction process (Bohn, 1997: 3-7):

- *Extract the data from the source system to intimidate schema*

  The extraction routines are developed to isolate only the data elements that will be needed in the data warehouse environment. The primary reason for the project team to transfer all data to an intermediate schema is to provide additional information to enhance data conditioning, cleaning and transformation routines.

- *Convert the intermediate schemas to load data*

  Once the data has been transferred to the intermediate schema, the project team will execute the conversion routines needed to clean and transform the data.

- *Aggregate the load data*

  In this phase the project team will sort the combined data based on predefined criteria. These criteria are usually developed based on common sense guidance and/or end user input in the event of more complex data elements. The aggregation of data occurs within the intermediate schema.

- *Migrate the load data from the staging area to the data warehouse server*

  Once the data is considered accurate, comprehensive and valid, the project team should relocate the data elements to the data warehouse server. This will be accomplished by loading the data in the database management system.

  In addition to ensuring that the data is successfully loaded on the server, the database management system will also ensure referential integrity by identifying offending records which should be corrected.

- *Validate the data*

  Validation of data does not only take place at the end of the extraction and conversion process, but is ongoing throughout the entire operation. Part of the validation program is the ongoing and integral involvement of the end user in the extraction and conversion process. The project team must therefore ensure that the end user is kept informed of any significant changes which could affect the final data presented.

  As part of validating the conversion of data, the project team will need to reconcile the data elements transferred from the various source systems to the

final data warehouse environment. High level reconciliations of this nature should occur at the following two stages and usually cover the total number of data records and any numeric fields which can be totaled:

i. After the source data has been extracted from the various source systems and transferred to the intermediate schema.
ii. After the source data has been aggregated and transferred from the intermediate schema to the data warehouse server.

4.9.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, the risks within the above mentioned process affects the effectiveness, reliability and integrity criteria of information.

The following detailed internal control risks are identified:

- The overall reliability of the final data elements could be affected by the incorrect transfer, loading and analysis of data prior to migration to the data warehouse environment.
- Unnecessary errors and the loss of data integrity could occur without an approved methodology for developing programmatic code for the extraction of data from source systems to the data warehouse.
- Non-existent conversion specifications could be responsible for certain data elements not being identified by the project team.

4.9.3 Internal control considerations

The following internal control considerations are applicable:

- The conversion specifications should be signed off by the project and technical user team before proceeding with the detailed migration.
- The programmatic code process should follow an approved methodology (similar to the process identified above).

- Discussions should be held with the end user as a means of identifying unnecessary data elements which should not be transferred to the data warehouse environment.
- Controls should be implemented to ensure that data elements are updated to reflect changes made to the associated records within the source system.
- During the data cleansing process, the project team should identify data element errors and have developed approved correction procedures.
- All transformation procedures used to convert data elements not adhering to the approved data codes should comply with the necessary data standards before being transferred to the data warehouse environment.
- The data elements should have been sorted based on an approved methodology before being transferred to the data warehouse environment.
- The project team should have taken steps to address offending records detected during referential integrity checks performed by the database management system.
- The reconciliation process performed as part of the validation procedure should take place at the two designated control points mentioned earlier in this phase.

5. Data warehouse package and vendor evaluation

5.1.1 Process steps

Identified as the second phase of the data warehouse development, the organisation will need to purchase an appropriate data warehouse application to access and analyse data (McManus, 1998: 1).

As part of this process, the project team, in conjunction with the organisation's procurement department, will need to consider two separate issues when deciding on which product to purchase, viz. vendor prescreening and the actual product selection criteria.

Vendor prescreening simplifies the vendor selection process and can save the organisation a significant amount of resources, in terms of time, costs, and human effort. This is accomplished by drastically reducing the number of comprehensive

vendor evaluations which would be performed for a first time encounter with a vendor (Tiwary S., Tewary A., 1998: 1). The questionnaire detailed under Annexure 4 provides a suggested framework which should can be applied in the evaluation of vendors and suitable applications.

To ensure the correct application choice, the project team should not only ensure the chosen data warehouse application meets existing user needs, but that it will also provide for the expected changes in user functionality in the foreseeable future (McManus, 1998: 1).

The project team, in conjunction with the procurement section, may opt to weight the various criteria based on their importance. If the selection committee do however decide to weight the selection criteria, input from the end user should be obtained in deciding the appropriate weighting (McManus, 1998: 4).

5.1.2 Internal control risks

According to COBIT's information criteria identified in chapter 1, the risks within the above mentioned process affects the effectiveness, availability, efficiency, and reliability criteria of information.

The following detailed internal control risks are identified:

- Incomplete and inaccurate vendor prescreening could result in a poorly supported data warehouse product being purchased.
- Unnecessary future costs can be avoided if a data warehouse application is purchased which is able to support future changes to the overall data warehouse environment.

5.1.3 Internal control considerations

The following internal control considerations are applicable:

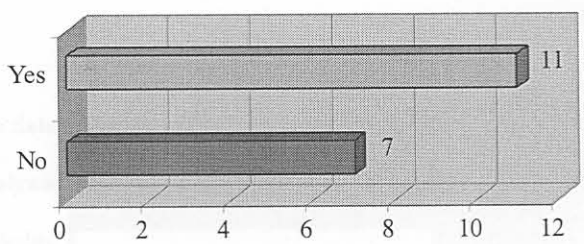
- The project team, in conjunction with the procurement section, should choose the most reputable vendor based on predetermined assessment criteria.

- Claims made by the supplier, such as Year 2000 compliance and financial stability of the vendor, should be supported.
- The selection committee should involve the end user as far as possible in the selection process.
- All decisions and comments relating to product selection should be documented and retained for future reference.

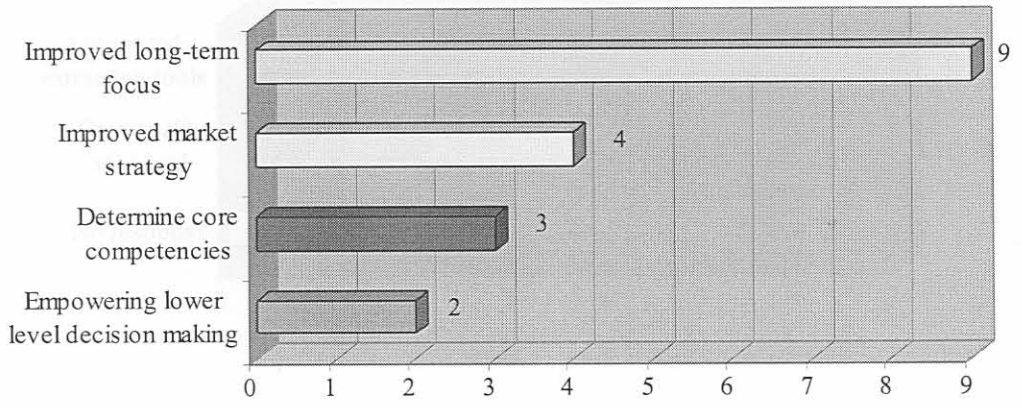
6. A South African perspective on the audit of developing data warehouse environments

As part of this study, a total of a 110 randomly selected internal audit heads of department were contacted regarding the internal control risks within the data warehouse environment. All of the 110 heads of department were registered with the South African Institute of Internal Auditors. A total of 18 replies were received (i.e. a 16% response rate) to the questionnaire sent (refer to annexure 1 for questionnaire). Results of the survey included in this section relate specifically to the development of the data warehouse environment:

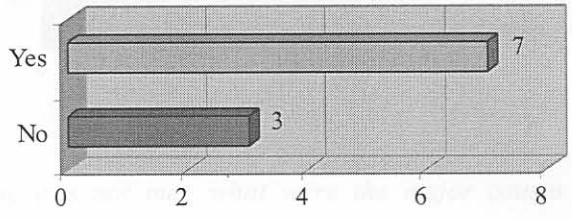
1. Does the company already have or is planning on implementing a data warehouse environment?



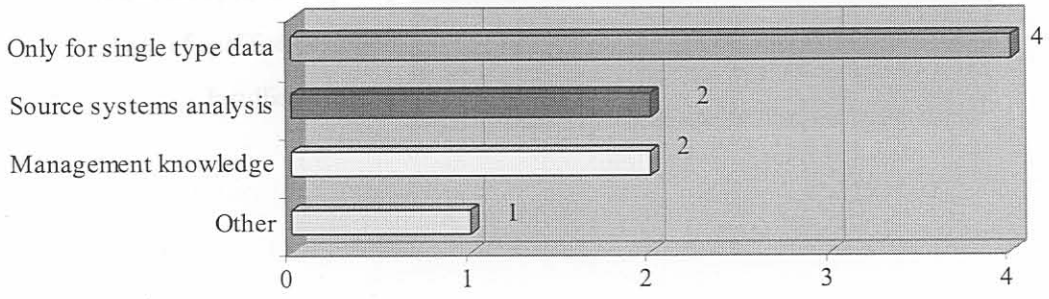
2. What was management's major intention in implementing the data warehouse?



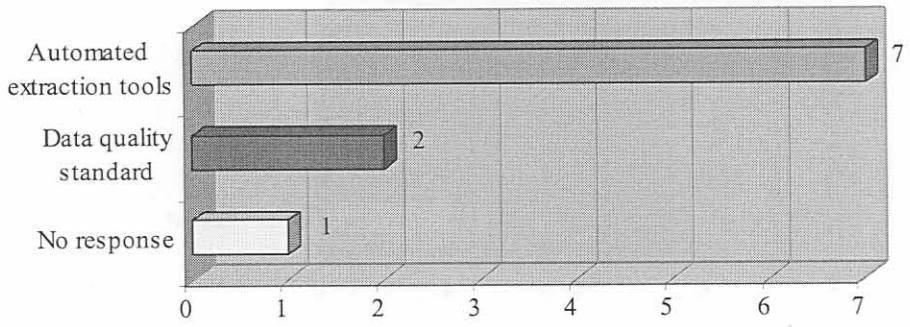
3. Did the Information Technology Department develop a system methodology specific for the data warehouse environment?



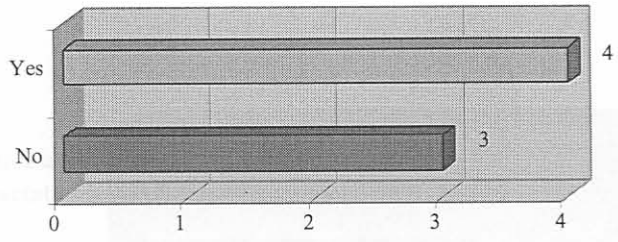
4. On what basis were all possible source systems identified?



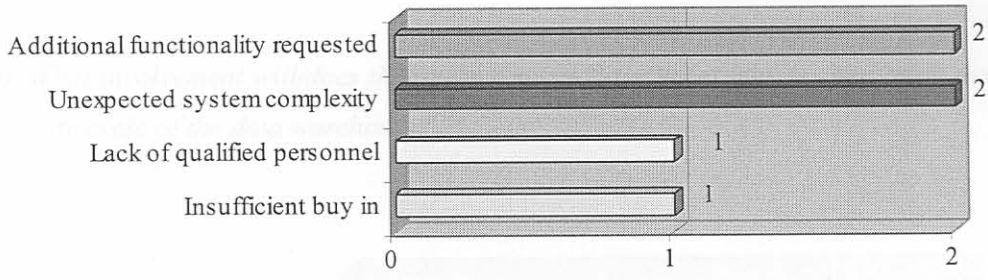
5. *What methodology was applied in ensuring that uniform data was introduced into the data warehouse?*



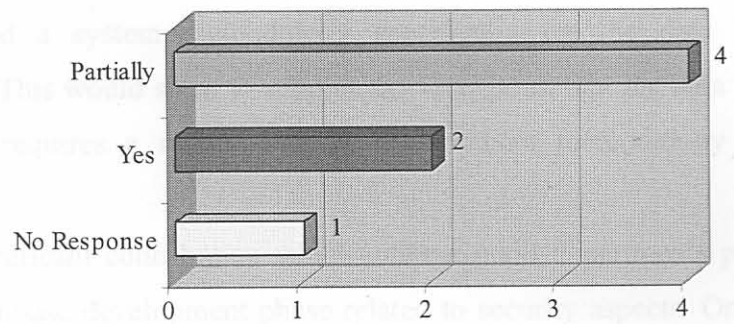
6. *Was the data warehouse implementation completed on time?*



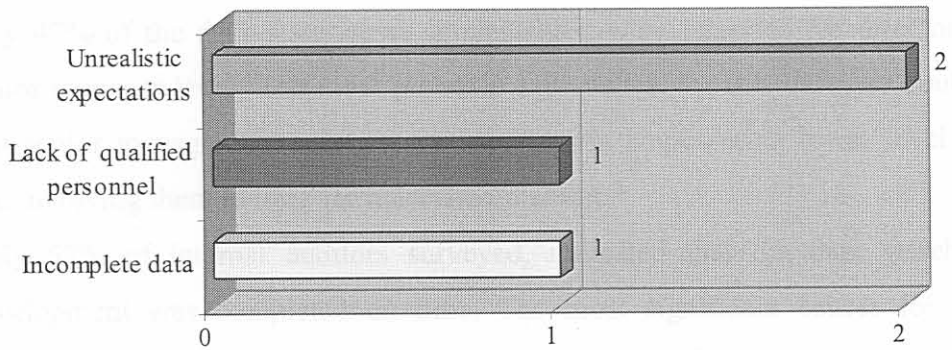
7. *If the deadline was not met, what were the major causes for the implementation not meeting the expected deadline?*



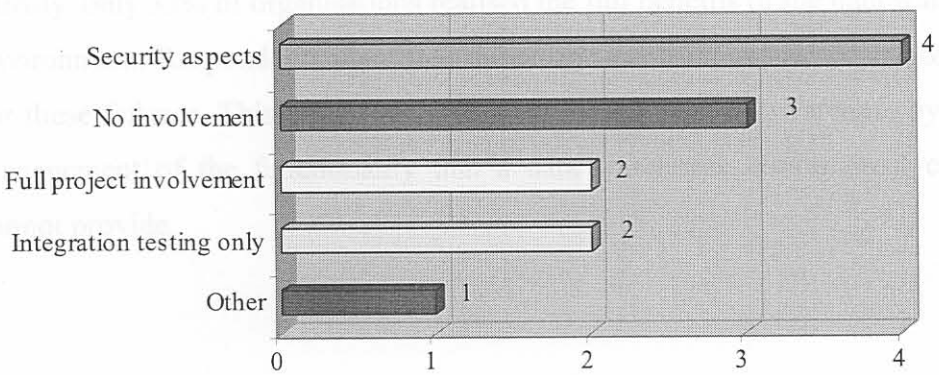
8. *Subsequent to the post implementation review, did management realise the expected benefits?*



9. *If the expected benefits were not fully realised, what were the major causes for management not realising these benefits?*



10. *What involvement will/does the internal audit department play in the system development life cycle of the data warehouse?*



Based on the above mentioned responses, the most significant findings raised included:

- Of the organisations who had embarked on a data warehouse development, 70% had developed a system methodology specifically for the data warehouse environment. This would seem to support the view point that the data warehouse development requires a unique system development methodology to ensure success.
- The most significant contribution which internal audit departments provided to the data warehouse development phase related to security aspects. Only 16% of audit departments indicated that they were involved in the entire project development. Possible causes of the lack of full involvement could be due to time constraints and the low level of criticality of the data warehouse environment as rated by various organisations.
- Although 50% of the respondents indicated that the primary intention of management in implementing a data warehouse was to improve long-term focus, only 47% of the data warehouses implemented were intended for director and senior manager level. The most probable explanation for this variance could be that senior management teams were focusing on empowering lower level staff and involving them in long-term decision making.
- Only 57% of internal auditors surveyed, indicated that the data warehouse development was completed on time. The most significant causes for these overruns were due to additional functionality requested by users and unexpected system complexity. Although not supported, it is probable that the majority of organisations who did complete their data warehouse developments on time, had not developed a system methodology specifically for the data warehouse environment.
- Finally, only 33% of organisations realised the full benefits of the data warehouse environment. Respondents identified unrealistic expectations as the major cause for these failures. This would seem to stem from a lack of awareness by senior management of the functionality that a data warehouse environment can and cannot provide.

7. Summary

In this chapter the reasons for the distinction between traditional system development life cycle models and those specific to the data warehouse were introduced. The study identified the possible internal control risks based on Inmon's system development life cycle for data warehouses. Suitable internal control considerations which could be used in assessing internal control risks were also provided.

In conclusion, South African trends relating to data warehouse developments were also provided.

8. Conclusion

Audit involvement in the development process is necessary to ensure control weaknesses are detected timeously and addressed with minimal resources. The system development life cycle for the data warehouse differs from that of traditional methodologies. Therefore the internal auditor should ensure that he/she is aware of the particular internal control risks which could exist in the environment. It is during the development phase that the internal auditor can and should contribute the most to a well controlled data warehouse.

The results of the empirical study supported the notion that a unique system development methodology for the data warehouse is required.

Chapter 3

Established Data Warehouse Environment

1. Introduction

The warehousing challenge is to technically capture, validate, integrate and transform data into meaningful information and then store that information into a data warehouse (Fryman, 1997: 46). In the preceding chapter we identified how the project team must develop suitable interfaces and integrate data warehouse applications needed to access data.

2. Aim

This chapter aims at identifying the internal control risks specific to the established data warehouse environment. The associated internal control considerations which may be applied in assessing the internal control risks are also discussed.

Results of the empirical survey conducted are provided at the end of the chapter.

3. Internal control risks and considerations within the established data warehouse environment

In this section we identify six internal control risks which may exist within an established data warehouse environment. Under each of the risks identified we provide a brief explanation of the risk and also indicate which of COBIT's information criteria, viz. effectiveness, integrity, availability, efficiency, confidentiality are affected.

The internal auditor is also provided with suitable internal control considerations which can be applied in assessing each of the internal control risks.

3.1 Inability to measure data quality and ensure satisfactory refreshing of data

3.1.1 Risk explanation

Without continuously monitoring data quality, management cannot ensure that data complies with approved management standards (Bohn, 1997: 1).

The refreshing of data within the data warehouse is fixed during the codification of the interfaces between the source system and data warehouse applications (Inmon, 1996: 280). If the refreshing rate of data it is not revisited with the user on a frequent basis, it is possible that such rates may become unsuitable in the future and result in users placing reliance on inaccurate data presented by the data warehouse.

According to COBIT's information criteria identified in chapter 1, the risk identified affects the effectiveness, integrity, availability and efficiency aspects of information (Curtis & Joshi, 1997: 40-43).

3.1.2 Internal control considerations

The following internal control considerations are applicable (ibid.):

- A data conversion plan should be developed.
- At a minimum, the data conversion plan addresses:
  i. The methodology applied in developing the data warehouse
  ii. Approved tolerance levels for errors in source data
  iii. The data standards and what data quality measures have been implemented.
- The management team should take steps to ensure that the transfer of data from the staging area to the final data warehouse is administered by the data administrator.
- All parties involved with the data warehouse should be familiar with the contents of the data conversion plan.
- The conversion plan should be updated with all new subject areas added to the environment.

- Tolerance levels for source data errors should be revisited regularly by the management team based on user feedback.
- Methods and monitoring procedures should be in place to assess the reliability and acceptability of data.
- Audit log issues recorded by data warehouses should be reviewed and significant anomalies followed up timeously.
- The time base applied in refreshing data should be determined based on a trade-off between timely data and the effective utilisation of information technology resources.

3.2 Not ensuring the completeness of data migrated to the data warehouse (Fryman, 1997: 46)

3.2.1 Risk explanation

In instances where management decides to include additional subject areas over time, ineffective project management and the lack of an approved development methodology will result in new subject areas not being included in the most efficient and effective manner.

Changes made to source systems without considering the data warehouse environment could affect the completeness of data migrated to such an environment (Inmon, 1996: 182). An ineffective communication process amongst the various Information Technology teams and end-users can result in these changes not being communicated effectively.

According to COBIT's information criteria identified in chapter 1, the risk identified affects the integrity aspect of information (Curtis & Joshi, 1997: 40-43).

3.2.2 Internal control considerations

The following internal control considerations are applicable (Fryman, 1997: 46):

- An approved data warehouse development methodology should be applied when adding new subject areas to the data warehouse environment (Inmon's development methodology detailed in chapter 2 is recommended).
- A proactive communication process should be in place to ensure that all changes made to source systems on which the data warehouse depends are reported timeously and to the correct personnel for action.

3.3 Ongoing availability of data warehouse operations cannot be ensured (Warigon, 1998: 55)

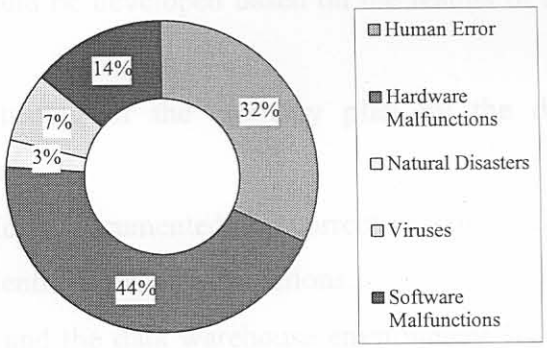
3.3.1 Risk explanation

A 1997 survey (Semer, 1998: 47) which tracked the major classifications of data loss among 50 000 organisations was conducted by Ontrack Data International Inc.. The survey indicated that 44% of data losses were caused by hardware or system malfunctions. Figure 3.1 provides more comprehensive results of the survey.

According to the Disaster Recovery Journal (Semer, 1998: 42), organisations suffered significant disaster-related costs in 1997:

- Each on-line outage averaged four hours and cost American companies and average of $329 000 in lost revenues and productivity.
- For each hour of unscheduled downtime, 355 worker hours were lost.
- Major businesses lost 38.1 million work hours, or $444 million in wages annually.

Figure 3.1 - Source of data loss



Source: Semer, 1998: 47

It is apparent from the above mentioned statistics that a significant risk is faced by organisations should mission critical systems become unavailable. Although the data

warehouse environment is only one source providing information to the user, the statistics provide an indication of the potential losses which can be incurred. Without consistent and supported data warehouse services, the end user may be unable to make informed management decisions.

According to COBIT's information criteria identified in chapter 1, the risk identified affects the availability aspect of information.

3.3.2 Internal control considerations

The following internal control considerations are applicable (ISACA, 1998):

- The data warehouse environment should be included in the business's overall continuity planning process.
- The data warehouse environment should be considered in the business impact analysis. This analysis identifies and inventories mission critical resources; quantifies the costs associated with failure to transact business due to the loss of resources; and estimates the downtime the organisation can bear while those resources are being restored.
- If the data warehouse environment is identified as one of the mission critical applications, management should identify and weight the risks specific to the data warehouse.
- Recovery plans and procedures should be developed based on the results of the initial business impact analysis.
- Routine training and simulation testing of the recovery plan by the data warehouse users should be conducted.
- Results of recovery plan tests should be documented and corrective action taken to address significant weaknesses identified during simulations.
- Interfaces between source systems and the data warehouse environment should be inventoried to ensure synchronisation of data during backup and recovery.
- Backup and recovery procedures should be fully documented, understood, accessible, enforced and tested regularly (Clark, Holloway & List: 114-115).

- The back-up and recovery procedures should take into account possible error conditions which could be encountered during the back-up and recovery process and provides suitable troubleshooting guidelines (ibid.)

3.4 *Overall data warehouse administration becomes ineffective and inefficient* (Warigon, 1998: 59)

3.4.1 *Risk explanation*

By not effectively monitoring the data warehouse environment, it can become unwieldy. In many instances, ineffective and inefficient data warehouses are caused by not regularly archiving outdated data and by not executing frequent capacity planning measures. The ultimate effect of not performing these activities are increased annual storage, processing and operating costs (ibid.).

Routine archiving of data involves the rolling up of outdated data to higher levels of summary (Inmon, 1996:69). This rolling of data can either be by means of transferring data from one level of the data warehouse architecture to another or, retaining data within a high-performance storage medium.

According to COBIT's information criteria identified in chapter 1, the risk identified affects the efficiency, effectiveness and availability aspects of information.

3.4.2 *Internal control considerations*

The following internal control considerations are applicable:

- Formalised assessments should be conducted in conjunction with the end user as a means of identifying data elements whose probability of access is close to zero (Inmon, 1996: 306). Factors which should be considered before data is archived (Inmon, 1998: 4):
  i. Time

     The project team must consider whether there is a probability that once archived that this data will be needed by the end user again. The costs of

restoring the data can sometimes exceed the cost of retaining it within the data warehouse environment.

ii. Classes of data

Ascertain what classes of data are most frequently used in queries and whether this pattern will change in the foreseeable future.

iii. Level of detail

Ascertain what level of detail is most commonly utilised by the end user and whether this pattern will change in the foreseeable future.

iv. Strategic importance of data

Although certain classes of data may not be accessed on a frequent basis, there is a probability that the class may be strategically important. In such instances, it is recommended that the data remain within the data warehouse environment.

- Maintenance and data management policies should be developed which clearly stipulate the methods and time frames which should be applied in phasing out unnecessary data classes (Zicker, 1998: 1).
- Management should consider scheduling and monitoring software which can simplify the tracking of outdated data and provide automated archiving functionality (ibid.).
- A data administrator should be employed or assigned the responsibility of monitoring and administering the archiving of data (Curtis & Joshi, 1997: 40-43).
- Statistics on performance, capacity, and availability should be provided (including historical versus forecast performance variance explanations) for the data warehouse on a regular basis (ISACA, 1998).
- Performance reporting information to users regarding usage and availability should take place (such reporting should include capacity, workload scheduling and trends) (ibid.).
- The package provider or data warehouse project team should be requested to give assurances that data warehouse applications will be able to manage growth in processing rates (Curtis & Joshi, 1997: 40-43).
- As part of the post-implementation phase of the system development life cycle, criteria should be included to determine the future growth and changes to performance expectations (ISACA, 1998).

3.5 *Data warehouse access is not restricted to authorised users* (Warigon, 1998: 55)

3.5.1 *Risk explanation*

Unauthorised access to data retained within the data warehouse can result in significant losses to the organisation (Warigon: 1998: 55-60). These threats can be caused by accidental or malicious attacks from employees. Outside threats can be caused by competitors. The result of such unauthorised access could be negative publicity for the organisation and a loss of continuity of data warehouse operations. Management will need to identifying security vulnerabilities which could negatively impact the organisation's image. As part of this assessment, physical security risks should also be considered. (ibid.).

According to COBIT's information criteria identified in chapter 1, the risk identified affects the confidentiality aspect of information.

3.5.2 *Internal control considerations*

The following internal control considerations are applicable:

- Management should classify data to ensure that the application of security resources is optimised and that different protective measures are used for different categories (Warigon: 1998: 55-60). Classifications of data could include public, moderately sensitive and highly sensitive data.
- Project management should quantify the value of data requiring protection. The criteria which could be used in determining the value of data requiring protection can include:
  i. The cost to reconstruct the data should a disaster occur.
  ii. The cost of restoring the integrity of violated data.
  iii. The inability to obtain data timeously thereby preventing informed decisions being made.

iv. Costs of litigation should customer's data be erroneously or intentionally exposed to unauthorised sources.

- Vulnerabilities to the data warehouse should be identified, evaluated and documented.
- A security policy should be developed. The policy should include (ISACA, 1998):
  i. Identification of security roles.
  ii. Security validation.
  iii. Documented proof of management support and commitment.
  iv. Access philosophy.
  v. Access authorisation procedures.
  vi. Annual reviews of access authorisation.
  vii. Password standards identified.
  viii. Security awareness drives.
  ix. The role of the security administrator defined.
- Confidentiality and intellectual rights agreements should be in place for all data warehouse users (ISACA, 1998).
- Each user should be defined to the database with a unique user identification.
- Passwords should be assigned to each user and the system pre-empts personnel to change theirs every thirty days.
- All access privileges should be approved by the data owner.
- Access rights should be revisited on a monthly basis to ensure that all terminated or transferred personnel are removed from the access rights to the system.
- Improved control is realised when management consider encrypting data. This should however only be considered in cases where data is extremely confidential. Encryption is costly and may prove cumbersome since the algorithms consume large portions of the central processing unit's resources (Warigon: 1998: 55-60).
- Ultimately, the management team should have selected the most cost effective security measures, i.e. the costs of protecting the data does not exceed the maximum monetary amount that the loss of data would represent (ISACA, 1998).

3.6 *Ongoing risk assessments over the data warehouse environment are not conducted* (Warigon, 1998: 55)

3.6.1 *Risk explanation*

Cost-effective measures are required to address the most significant risks within the data warehouse environment. Organisations should be focusing on ways to limit costs and only secure mission critical assets (ibid.). This valuable information can only be obtained by performing, and revisiting risk assessments over information technology environments such as the data warehouse. These assessments will allow management to identify how critical the risks are within the data warehouse environment and thereby apply the limited resources in the most effective and efficient manner.

Ultimately, if organisations do not perform risk assessments on a frequent basis, the effective and efficient utilisation of resources cannot be ensured.
According to COBIT's information criteria identified in chapter 1, the risk identified affects the effectiveness, integrity, efficiency and reliability aspects of information.

3.6.2 *Internal control considerations*

The following internal control considerations are applicable (ISACA, 1998):

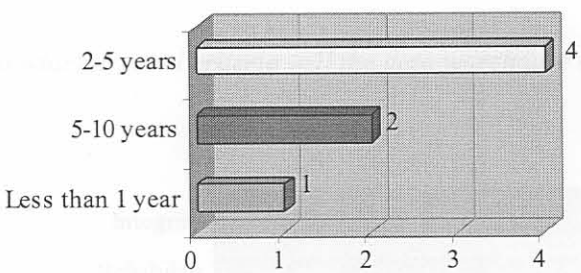
- Management should have a program in place to identify additional data types which could provide additional benefit to the user if migrated to the data warehouse environment.
- Overall warehouse administration should meet user's expectations (e.g. conduct user satisfaction surveys).
- Regular benchmarking with similar facilities should take place to ensure best practice.
- Data quality control should be revisited to ensure that the number of instances of contaminated data going undetected is reduced.
- Company wide business reviews should be performed to identify, investigate and resolve data elements that are not within quality standards.

- Mechanisms should be in place to record and monitor the data warehouse's capturing of data and the quality of such data over time.
- Frequent security assessments should be undertaken (Warigon, 1998: 60). Evaluations should be conducted continuously to determine whether security measures and controls are:
  i. Simplistic and straightforward.
  ii. Carefully monitored.
  iii. Do not hamper authorised users from performing their duties effectively and efficiently.
  iv. Are easily adaptable to necessary changes in control standards.

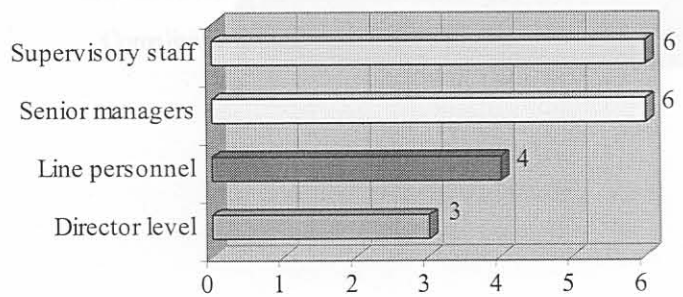
4. A South African perspective on the audit of established data warehouse environments

The results of the local survey are featured below. The results relate specifically to the internal control risks within the established data warehouse environment:

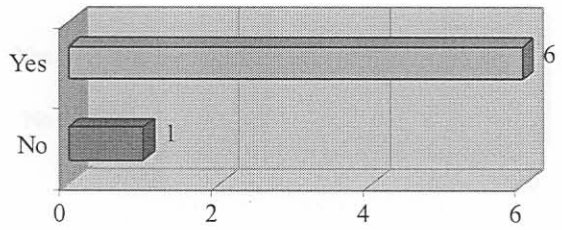
1. How long has the organisation had a data warehouse?



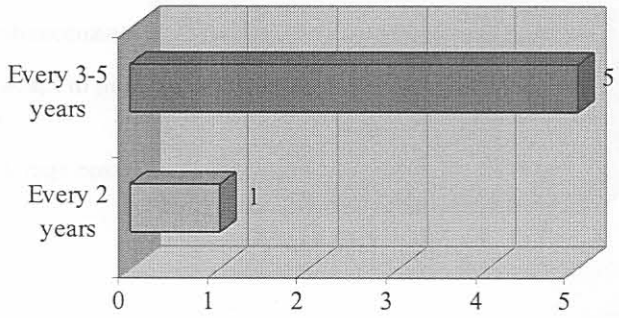
2. Which level of staff utilise the data warehouse structure?



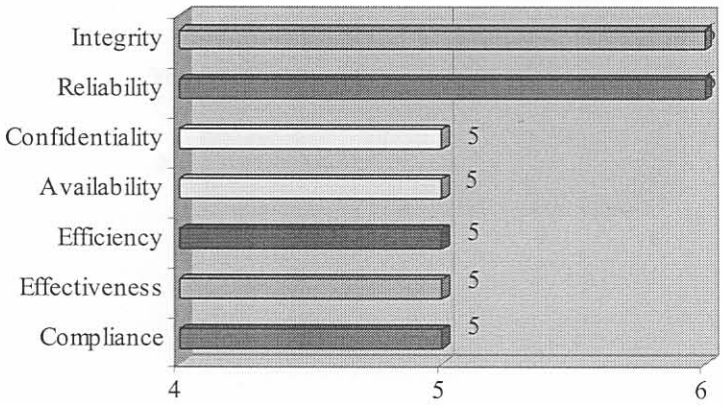
3. *Is the data warehouse environment identified as an application reviewed by the internal audit team on a periodic basis?*



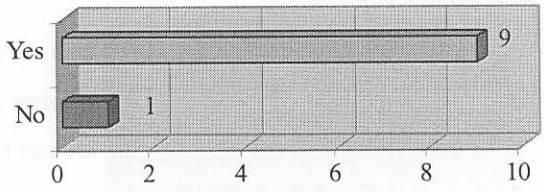
4. *If it is audited, how frequently will the data warehouse environment be reviewed by internal audit?*



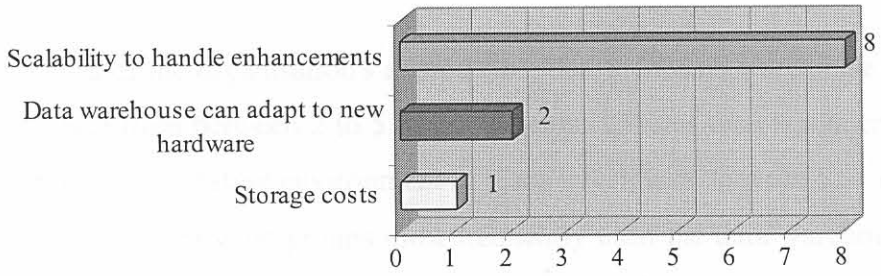
5. *According to which control criteria will the data warehouse environment will be reviewed?*



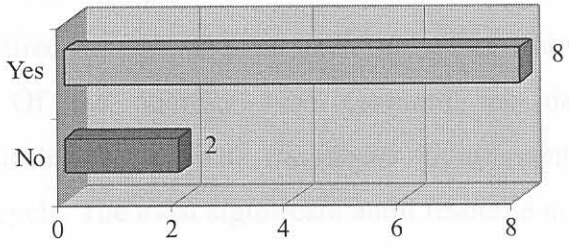
6. Was capacity management identified as part of the audit approach?



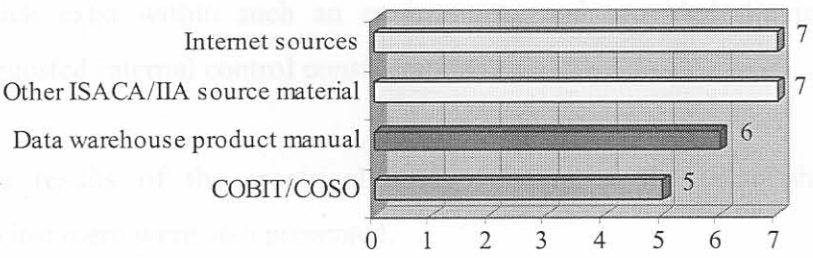
7. If capacity planning was considered, for which of the following reasons was it considered?



8. Will/has the data warehouse environment been included in the organisation's continuity plans/efforts?



9. *Which audit resource materials were used in formulating a suitable audit approach and program?*



Based on the above mentioned responses, the most significant findings raised included:

- 57% of the organisation's approached, had their data warehouse environments in place from between 2 to 5 years. Whereas 25% of the respondents indicated that they had a mature environment in place which was between 5 to 10 years old.
- When asked what groups most frequently used the data warehouse environment, 63% of the respondents indicated that the data warehouse was applied by senior management and supervisory staff. Only 15% indicated that the data warehouse environment was being used at director level. This strongly indicates that localised data warehouses are focused on middle management level (i.e. medium term planning) as opposed to the strategists within each of the organisations.
- Internal auditors stated that the data warehouse environment was an application which had been identified as part of their audit universe. Only 1 respondent indicated otherwise. Of the auditors who assessed the data warehouse environment, 83% indicated that the data warehouse environment was reviewed on a 3 to 5 year audit cycle. The most significant audit resource materials used in preparing a suitable audit approach were obtained from internet sources and other professional body materials.

5. Summary

In this chapter we identified the internal control risks which may exist within the established data warehouse environment. We also identified six internal control risks which exist within such an environment and provided the internal auditor with suggested internal control considerations.

The results of the empirical study relating to the established data warehouse environment were also presented.

6. Conclusion

Significant internal control risks are not only found in an environment where a data warehouse is being developed but also in an established data warehouse environment.

The internal auditor's primary aim is to ensure that management attain their primary goals and objectives. He/she should therefore ensure that data warehouse reviews are all founded both on management's assessment of how critical the data warehouse environment is and the risk assessment model adopted by the internal audit department. This approach will ensure that the internal auditor evaluates the data warehouse on a basis commensurate with the overall risk the organisation could be exposed to.

In reviewing the results of the empirical study, we can say that the most significant audit resources used in preparing the internal auditor's approach were obtained from internet resources. It appears that the data warehouse environment is in actual fact a relatively new audit cycle with limited resource materials available to the internal auditor.

Chapter 4

Dependant Data Mart Environment

1. Introduction

In chapter 3 we described an organisation wide data warehouse. Although these types of data warehouses were the first to be introduced, they can sometimes fail to provide information to the end user effectively and within an acceptable time frame (Bersin, 1996). To address these weaknesses, an organisation may choose both an organisation wide data warehouse that addresses all operations and a simplified data structure for selected business units or subjects.

The following are primary reasons why such an environment is more favourable than a single organisation wide data warehouse (D2K, 1996):

- Business departments which implement a data mart as an extension of the data warehouse are able to customise the flow of data introduced into the data warehouse.
- The amount of historical data needed by the business function will be far less. Accordingly, the business unit will have access to historical data relevant only to their operations. This in turn will improve access and query response times.
- The business function will be able to select data extraction and analysis tools specific to their needs instead of being required to use an organisation wide package which might not cater for their exact needs.
- The unit cost of processing and storage on the size machine that is appropriate to the data mart is significantly less than the unit cost of processing and storage for the data warehouse environment.

The key advantage of such a data structure is improved response times to access data and process queries (Bersin, 1996).

2. Aim

In this chapter we provide the internal auditor with an understanding of how the dependant data mart relies on the existing data warehouse. We also aim to show what control risks exist within this environment. We will however only identify the outstanding internal control risks which are unique to the data mart environment in this chapter. This is because certain internal control risks and considerations identified in chapter 2 and 3 of this study also apply to the data mart environment.

This chapter concludes by providing suitable internal control considerations which can be applied in assessing these unique internal control risks.

3. Understanding the dependant data mart environment

3.1 Background

A data mart is generally defined as a database or collection of databases designed to help managers make strategic decisions about their business. A data warehouse however combines databases across an entire organisation. Data marts are usually smaller and focus on a particular subject or department (AOL, 1996).

3.2 Development considerations

The development of the data mart environment should comply with a development methodology similar to that outlined in Chapter 2. Since the scope of the development project will be restricted to a single business unit or subject, the project will need to focus more aggressively on efficiency and effectiveness. This will ensure that project deliverables are attained within a suitable timeframe and within budget constraints.

The data mart's key advantage over a single organisation wide data warehouse is improved response time. Therefore increased focus should be applied in ensuring rapid response to queries raised by users. To assist in achieving the goal of improved response time, the development team should adopt a more rigid development strategy

than that adopted for the organisation wide data warehouse. Accordingly, this strategy should consider the following specifications (Bersin, 1996):

- The most effective extraction and interrogation tools should be selected to ensure that queries are processed in the shortest possible time frame.
- Data marts should be able to address an increase in capacity over time. Studies show that data marts grow in size by 30 to 100% per annum if the full functionality of the data mart is realised by the user.
- Common data elements should be used if the organisation intends on implementing a number of dependant data marts for certain business units or subjects. This will ensure that data is consistent in quality.
- Ongoing monitoring of the data mart environment is in place where the size of the environment dictates.

The transfer of data from the existing data warehouse to the data mart should be controlled. This interface between the two environments should consider (D2K, 1996):

- The interface should execute on a periodic basis.
- The interface may update the environment only with new data or may refresh all data resident within the environment.
- The interface should link the metadata resident within the data mart to that resident within the originating data warehouse. The interface needs to describe how the data between the two environments is linked so that if the user drills down between the two environments, the data marts meta data can easily find the heritage of the data within the data warehouse.

4. Internal control risks and considerations within the data mart environment

In this section two unique internal control risks are identified which may exist within the data mart environment. Under each of the risks identified a brief explanation of the risk is provided. We will also indicate which of COBIT's information criteria, viz. effectiveness, integrity, availability, efficiency, confidentiality are affected.

The internal auditor is provided with suitable internal control considerations which can be applied in assessing each of the internal control risks.

4.1 A lack of sufficient response time monitoring on a periodic basis

4.1.1 Risk explanation

The inconsistent or total lack of ongoing response time monitoring within the data mart environment may result in information not being provided to the user on time. In such instances the user will either stop utilising the data mart environment or be unable to make informed management decisions (Bersin, 1996).

According to COBIT's information criteria identified in chapter 1, the risk identified affects the efficiency, effectiveness and availability aspects of information.

4.1.2 Internal control considerations

The following internal control considerations are applicable (Bersin, 1996):

- As part of the monitoring procedures, the management team should monitor what data is being accessed, what response times are being achieved and how much data is requested. They will also ascertain what the busiest times of the day, week and month are.
- Escalation procedures should be in place to ensure that unsatisfactory trends are reported upon and actioned.
- Users should be trained on what is considered acceptable response times and what to do if the data mart environment does not meet minimum standards.
- Controls should be in place to limit the number of ad-hoc queries which are not catered for as part of the pre-prepared reports. They should also identify queries during the data mart's initial project development.
- The network on which the data mart relies upon should be monitored on a periodic basis to ensure that no related network hardware or software problems exist.

4.2 Transfer of data from the organisation wide data warehouse to the data mart is not controlled

4.2.1 Risk explanation

The data mart may become inefficient and ineffective if the project team fails to ensure that only the most necessary data is transferred from the organisation wide data warehouse to the data mart (Bersin, 1996).

The upload of data from the data warehouse to the data mart may also create data integrity and availability problems if not controlled. The uncontrolled uploading of data may arise if:

- The frequency of updating data is not in agreement with the end user's needs.
- The data mart is not updated with data changes made in the data warehouse environment.
- The inability of the data mart environment to notify the users of subsequent changes made to data already relied upon, may result in incorrect management decisions being made. This is applicable in instances where a total refresh of data occurs.

In all three instances, the significant loss of decision making and the reliance on incorrect data can cause significant financial losses to the organisation.

According to COBIT's information criteria identified in chapter 1, the risks identified affect the availability, integrity and reliability aspects of information.

4.2.2 Internal control considerations

The following internal control considerations are applicable (Bersin, 1996):

- Procedures should be in place to ensure that the data transferred to the data mart is based upon assessments performed which determine which data types are most frequently accessed by the end users within the originating data warehouse. This

will guarantee that the data mart's response times are efficient and that only the most needed data is retained within the data mart environment.

- Procedures should be in place to ensure that all data transferred from the operational system is transferred at the most appropriate time. The project team in conjunction with the end user should identify the time when the upload of data should take place so as to reflect the correct data characteristics.
- Users in conjunction with the project team should have agreed upon whether the uploading of data to the data mart should be on a total or partial refresh basis.
- The project team should have taken steps to ensure that the procedures which will be followed in updating data already resident within the data mart complies with an approved methodology.
- Suitable checks should be in place to ensure that any significant changes made to operational data already relied upon by the user are effectively followed-up and communicated.

5. Summary

The dependant data mart is an extension of the data warehouse. This extension allows for improved access to data by segregating data types according to business functions or subjects.

In addition to the internal control risks identified in chapters 2 and 3, we identified an additional two internal control risks relating to the dependant data mart. These are:

- A lack of sufficient response time monitoring on a periodic basis.
- The transfer of data from the organisation wide data warehouse to the data mart is not controlled.

Suitable internal control considerations were also provided as a means of assessing the extent of these internal control risks.

6. Conclusion

For organisations seeking improved response times from their organisation data warehouses, data marts are fast and inexpensive to implement and are able to show a

fast return on investment. To achieve this however, it is important to realise that the data mart should be developed according to a rigid development methodology, similar to that introduced in chapter 2 of this study.

To ensure that the data mart environment also provides ongoing benefit to the organisation, the environment should be monitored on an ongoing basis to ensure that it does not become unmanageable over time.

Chapter 5

Distributed Data Warehouse Environment

1. Introduction

There exist a number of reasons why a centralised data warehouse is considered to be the most appropriate data structure (Inmon, 1996: 197):

- Data distributed across multiple locations is usually cumbersome to access.
- The volume of data in many instances, necessitates a centralised data warehouse repository.
- It is only at the centralised processing operation that an integrated view of data will add the greatest value.

Although the above mentioned factors substantiate the creation of a centralised data warehouse, there exist certain instances where a distributed data warehouse will be more appropriate (Bell, 1992: 2-4):

- Experience has shown that 90% of data operations are local, meaning that in instances where organisations are dispersed geographically, the need for users to access their data locally is increased.
- For back-up purposes, it is considered good business practice to have data replicated in a number of sites to ensure continuos operation.
- Improved technology addresses limited access to data by centralised processing operations.
- The ability of distributed data warehouses to be expanded due to increasing data volumes is easier than that of a corporate data warehouse (Inmon, 1996: 213).

2. Aim

In this chapter we provide the internal auditor with an understanding of the distributed data warehouse environment. We also identify what internal control risks should be

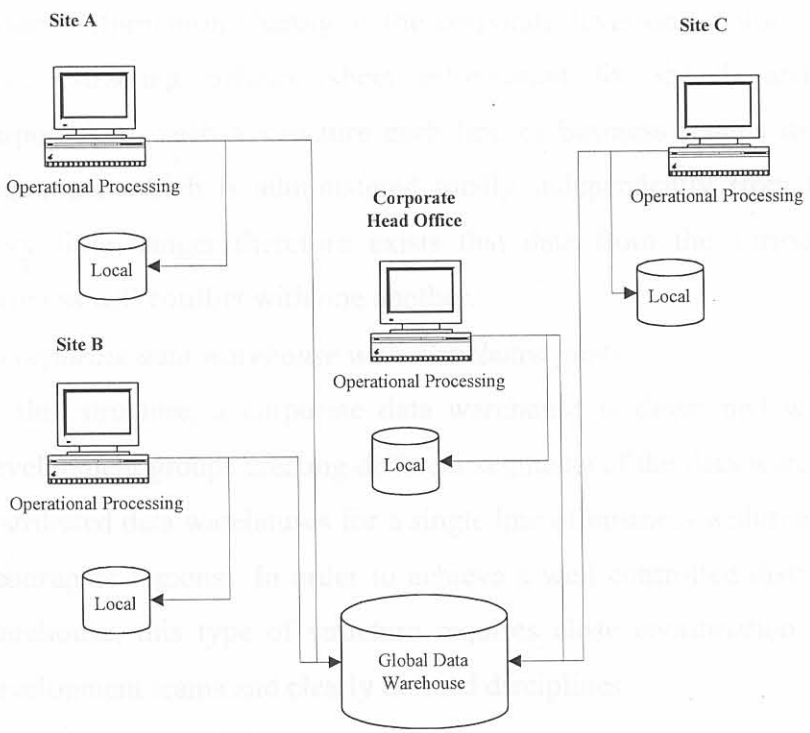
considered during an evaluation of such an environment. This chapter concludes by providing suitable internal control considerations which can be used in assessing internal control risk. The result of the empirical survey conducted as part of the study is also included.

3. Understanding the distributed data warehouse environment

3.1 Background

In describing the distributed data warehouse environment, Inmon differentiates between a local and global data warehouse structure (Inmon, 1996: 200-202). The local data warehouse structure incorporates all data required for decision making at the various sites. Whereas the global data warehouse has as its scope, data relating to all business units within the organisation (Inmon, 1996: 202). Figure 5.1 provides an example of a distributed data warehouse environment. From the figure, we can see that no links exist between the various local data warehouse structures and the global data warehouse.

Figure 5.1 - The global data warehouse structure



Source: Inmon, 1996: 205

All data transferred to the global data warehouse is derived directly from the various source systems located at each of the organisation's sites. Inmon indicates (Inmon, 1996: 205) that the data structure for the global warehouse is designed and defined centrally. He says it is also necessary to identify what data will be retained within the global warehouse. This decision is taken by the local designer and developer in conjunction with the corporate management team. The primary reason for this, is that data retained within the local and global data warehouse is mutually exclusive. This means that data located within the local data warehouse should not also be included in the global data warehouse (Inmon, 1996: 209).

3.2 Development considerations

The development methodology applied in establishing a distributed data warehouse should comply with the process detailed under chapter 2. However, the integration of distributed data warehouses can assume either of three specialised structures (Inmon, 1996:215-217):

- *Separate and un-integrated lines of business*
  Although considered a rare structure, it is possible to find a situation where the organisation has totally unrelated lines of business which will only require information sharing at the corporate level on an infrequent basis (e.g. extracting balance sheet information for month and year-end purposes). In such a structure each line of business retains its own data warehouse which is administered totally independently from the others. Very little danger therefore exists that data from the various lines of business will conflict with one another.
- *A corporate data warehouse with distributed parts*
  In this structure, a corporate data warehouse is developed with various development groups creating different segments of the data warehouse (e.g. distributed data warehouses for a single line of business within a number of geographic regions). In order to achieve a well controlled distributed data warehouse, this type of structure requires close coordination among the development teams and clearly defined disciplines.

- *Various levels of data within a single corporate data warehouse*

  This structure is also considered more common, and far easier to manage than the corporate data warehouse with distributed parts. In this instance the various project teams develop the diverse levels of data based on the predetermined levels of granularity required (e.g. summarised data, detailed data, etc.).

3.3 *Access and security considerations*

In order to ensure confidentiality and completeness of transfer, the management team will need to consider how the information transferred to and from the global data warehouse to the various sites is protected.

Three factors exist within the distributed data warehouse environment which can ensure more secure access to data. These factors are:

- *Policy standards*

  In the distributed data warehouse environment, management may opt for various security policies as an additional means of addressing security concerns (Bell, 1992: 283-284). The selection criteria used in selecting the most appropriate security policy will depend on management's assessment of access and data transfer risks.

- *Identification and authentication*

  The major portion of current day distributed data warehouses allow users access to global data via their local sites. This is done by allowing the local site to perform the user identification and password verification. Once the user has been admitted by the local site, all other sites will accept user requests within the limitations of the predetermined access rights (Bell, 1992: 293).

- *Encryption*

  In terms of ensuring safe transfer of data from one site to another, the most well known protection routine is encryption (ibid.). The use of the internet has proliferated and the ability of remote users to access the systems has also increased. Improved controls are needed to remedy this new class of weakness. Encryption of data allows for data to be transported across an unsafe network.

If the data is intercepted by a perpetuator it will bear no meaning because of the coding structure applied to the original transmission. The intended receiver will be able to process the data based on the fact that he or she is in possession of the decryption key which will translate the message into intelligible data (Inmon, 1997: 10).

The database or data warehouse encryption process has the following unique characteristics:

- The primary principle of this technique is to ensure the internal structure of data being transferred is consistent. This means, that the sender's message must be encrypted and decrypted exactly into the same length field – no variations allowed.
- This principle also applies to the format of the message being sent. If the message is ASCII format, it must be encrypted and decrypted as such.
- Data must be stored in an intermediate location, such as a database.
- Database or data warehouse encryption must also allow for split encryption. This ensures that the database administrator can encrypt only certain sections of the data being transmitted.
- Interleaved encryption must also be provided for. This process allows the database administrator to encrypt selections of data with different algorithms or keys.

4. Internal control risks and considerations within the distributed data warehouse environment

In this section we identify three unique internal control risks which may exist within the distributed data warehouse environment. Under each of the risks identified we provide a brief explanation of the risk and also indicate which of COBIT's information criteria, viz. effectiveness, integrity, availability, efficiency, confidentiality are affected.

The internal auditor is also provided with suitable internal control considerations which can be applied in assessing each of the internal control risks.

The internal auditor's review of the distributed data warehouse environment should include the principal internal control considerations identified in either chapter 2 or 3 of this study as well as the considerations below. The applicability of the internal control considerations identified in chapter 2 or 3 will depend on whether internal audit is involved during the development phase of the data warehouse or whether an assessment of an established system is being performed.

4.1 Distributed data warehouse access is not restricted to authorised users

4.1.1 Risk explanation

There are two major risks concerning unrestricted access to the distributed data warehouse environment (Bell, 1992: 5):

- Ensuring controlled access across open communication channels.
- Ensuring optimal access to distributed resources.

Significant risk of unauthorised disclosure of information may occur if access restrictions do not ensure that only valid users have rights to view data retained within the data warehouse environment.

An optimal access consideration involves the risk of inefficient access to data. This could result in users under utilising the data warehouse due to poor response times.

According to COBIT's information criteria identified in chapter 1, the risk identified affects the confidentiality, integrity and efficiency aspects of information.

4.1.2 Internal control considerations

The following internal control considerations are applicable (ibid.):

- The company-wide security policy addresses the specifics of the distributed data warehouse environment.

- Authorisation rules used to gain access to the global data warehouse are retained at each of the local sites.
- Suitable controls are in place to ensure that authorisation rules are in line with the access philosophy approved by the organisation's senior management team.
- The database or warehouse encryption technique is being applied during the transmission of critical data elements across open communication channels.
- Mechanisms are in place to monitor access performance and monitored results are reviewed for areas of concern on a frequent basis. The utilisation of benchmarking to gauge results against other leading organisations should have been considered.
- All distributed data warehouses are resident behind a well controlled firewall and all communications are processed through the firewall. The following properties relating to the firewall should be in place (ISACA, 1998):
  i. All traffic from inside to outside, and vice-versa, passes through the firewall (this should not be limited to logical controls, but should also be physically enforced).
  ii. Only authorised traffic, as defined by the local security policy, should be allowed to pass through the firewall.
  iii. The firewall itself is immune to penetration.
  iv. Traffic is exchanged through the firewall at the application layer only.
  v. The firewall architecture combines control measures both at the application and network level.
  vi. The firewall architecture enforces a protocol discontinuity at the transportation layer.
  vii. The firewall architecture deploys strong authentication for management of its components.
  viii. The firewall architecture hides the structure of the internal network.
  ix. The firewall architecture provides an audit trail of all communications to or through the firewall system and will generate alarms when suspicious activity is detected.
  x. The firewall architecture defends itself from direct attack (e.g. through active monitoring of traffic and pattern recognition technology).

4.2 Ongoing availability of the distributed data warehouse operations cannot be ensured

4.2.1 Risk explanation

Similar to the risk for an established data warehouse environment identified in chapter 3, the distributed data warehouse is also prone to expected and unexpected failure (Bell, 1992: 5). The most common forms of failure within the distributed data warehouse environment can be summarised as follows (Bell, 1992: 233-239):

- *Local transaction failures*

  These failures are caused either by unforeseen transaction failures, (such as system logic errors), or by system induced failures, (such as management override of computer programs or the intentional shut-down of computer operations). The severity of these failures are usually limited since they only affect a small number of transactions.

- *Site failures*

  Sites operate independently in the distributed data warehouse environment. Therefore it is possible for certain sites to be operational while others have failed (referred to as partial failures). Partial failures are considered far more hazardous than a complete failure of the distributed environment. This is because it is difficult for other sites to detect instances where other reliant sites are unavailable.

- *Media failures*

  Media failures are caused by hardware corruptions. The most common of these failures occur in hard disk storage devices.

- *Network failures*

  Networks are considered to be the back-bone structure used to ensure efficient and effective communications between the local and global sites. Although today's networks are considered to be robust, it is possible that line failures may corrupt communications. To a large degree, the ability for system software to reroute communications has overcome this failure type.

According to COBIT's information criteria identified in chapter 1, the risk identified affects the availability aspect of information.

4.2.2 Internal control considerations

The following internal control considerations are applicable (Bell, 1992: 233-239):

- Management should consider the effects of local transaction, site, media and network failures on the distributed data warehouse environment.
- The effects of the above mentioned failures should be quantified and suitable continuity plans developed to address the significant and controllable weaknesses identified.
- A suitable administrator should be appointed to ensure the regular updating of the documented continuity plans.
- Consistent testing of the distributed continuity plan should be performed and results of testing documented. Significant shortfalls identified during testing should be highlighted and addressed timeously by the management team (ISACA, 1998).

4.3 Efficiency of processing within the distributed data warehouse is not maximised

4.3.1 Risk explanation

The efficiency of transformation and integration of data between distributed sites is identified as one of the major risk areas in a distributed data warehouse environment (Bell, 1992: 5). If control mechanisms are not implemented to mitigate this risk, it is possible that users may become disillusioned when the provision of data warehouse functionality is slowed. This under utilisation of assets could result in uniformed management decisions been taken which could in turn affect the profitability and even the continued operation of the organisation.

Query optimisers can be utilised to address the controllable weaknesses relating to the efficient transfer and communication of information in a distributed data warehouse environment (Bell, 1992: 124). The task of the query optimiser is to govern and

expedite the processing and data transmission required for responding to queries. It in turn ensures that either the total cost or the total response time for a query is minimised (ibid.).

The optimiser operates by taking the user's query and applies four processing steps (Bell, 1992: 123):

- Determine the order in the which the various elements of the user's query should be executed.
- Identify the most suitable method available to access the required units of data.
- Identify the suitable algorithms needed to carry out the operation (this usually involves the program logic required to collate and order the data into meaningful and accurate results for the user).
- Identify the order that should be followed for the data movements between the various affected sites.

Two different forms of query optimisers can be utilised by an organisation. The selection of the most appropriate type of optimiser depends on whether the organisation wishes to realise savings in terms of costs or in time execution (Bell, 1992: 124-130):

- *Execution cost optimisers*

  This optimiser is used to minimise the use of the total system resources for a query and thereby reduce total operation costs.
- *Response time optimisers*

  The primary aim of this optimiser is to reduce response time rather than the total cost of the query processed by the user. The optimiser operates on the basis of determining the critical path needed to gain the necessary information in the shortest possible time.

The topic of query optimisation is a highly technical and complex one (Bell, 1992: 122-123). The internal auditor's overriding concern is to ensure that management are aware of the availability of these tools and that where feasible, the utilisation of the most appropriate optimiser has been considered. If not considered, the negative

impact on the efficiency and effectiveness of user queries within the distributed data warehouse environment can be significant.

According to COBIT's information criteria identified in chapter 1, the risk identified affects the effectiveness, availability, efficiency and the reliability aspects of information.

4.3.2 Internal control considerations

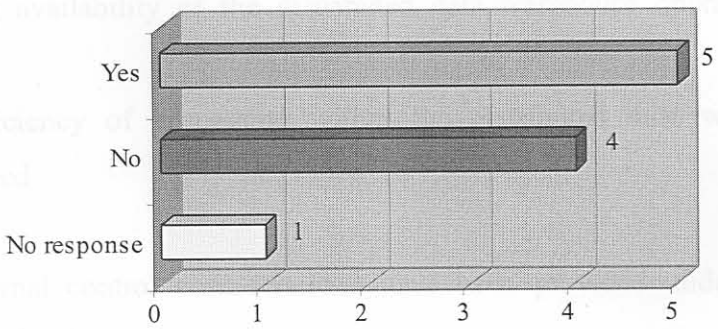
The following internal control considerations are applicable (ibid.):

- Query optimisers should be assessed by management as a means of improving processing response times or reducing total processing costs.
- The most appropriate query optimiser should be implemented, based on the needs of the end user, i.e. either an execution cost or response time optimiser.
- An approved service level agreement should be in place between the various user departments and the management team responsible for administering the distributed data warehouse environment.
- If a service level agreement is in place, the following minimum areas should be addressed as part of the agreement relating to the distributed data warehouse environment (ISACA, 1998):
  i. Quantified availability ratios.
  ii. Guarantees on the reliability of service from the supporting management team.
  iii. Levels of support which should be provided to users.
  iv. Capacity for growth and how frequent this issue should be revisited by the management team.
  v. Minimum acceptable level of service.
  vi. Details on the service charges for maintaining the distributed data warehouse environment.

5. A South African perspective on the distributed data warehouse environment

The results of a local survey conducted are as follows. They relate specifically to the internal control risks within the distributed data warehouse environment:

1. Will/is your organisation's data warehouse environment distributed in nature?



50% of the local internal auditors indicated that their data warehouse developments were distributed in nature. This indicates that a fair number of local organisations are utilising this unique structure. It is therefore imperative that the internal auditor become familiar with the risks pertinent within such an environment so that he/she is able to provide the necessary audit assurances to senior management.

6. Summary

In this chapter we introduced the distributed data warehouse environment and how this structure differs from the centralised data warehouse environment. The most significant differences between the centralised and distributed data warehouse environment are:

- Data elements are located at various dispersed sites. With a distributed global data warehouse housing data relating to all business units is located at the corporate site.
- Data elements loaded into the distributed global data warehouse are derived from the various operational systems located at the dispersed sites.
- The structure and content of the distributed global data warehouse is decided at the corporate site.
- The mapping of data into the distributed global data warehouse located at the corporate site is decided at the various dispersed sites.

The three unique internal control risks identified within the distributed data warehouse environment are:

- Distributed data warehouse access is not restricted to authorised users.
- Ongoing availability of the distributed data warehouse operations cannot be ensured.
- The efficiency of processing within the distributed data warehouse is not maximised.

Suitable internal control considerations have been provided under each of these internal control risks.

7. Conclusion

The distributed data warehouse environment has its difficulties. These are optimising the evaluation of queries, controlling access to the various data elements and ensuring the ongoing availability of data warehouse operations.

Half of the local internal auditors surveyed indicated that their data warehouse environments were distributed in nature. It would therefore seem imperative that a considerable amount of attention should be given to the impact of this unique structure on internal control risk. Although the impact of the distributed data warehouse environment on the internal auditor's assessment of internal control risk was considered, it is important that the internal auditor consider these factors in conjunction with those relating to the established data warehouse. It is only by combining these internal control considerations that the internal auditor will be able to perform a comprehensive evaluation of the data warehouse environment.

Chapter 6

Future Developments and Trends

1. Introduction

Data warehousing is a form of technology identified as one of the desired applications needed by organisations today and in the future (Du Plessis, 1998: 1). It is important that mangement and the internal auditor are aware of expected developments in the data warehouse environment. This will enable them to manage internal control risks.

Another significant future trend is the increase in utilisation of data warehousing technology by internal auditors. Any tool that allows the internal auditor to identify risks more effectively can result in the internal audit department being more efficient (Geiger, 1997: 31). Internal auditors use data warehousing as a means of assessing internal controls within other audit cycles. Routine audit cycles may include: accounts payable and receivable reviews, marketing and advertising assessments, fraud investigations, etc.

2. Aim

This chapter identifies what future data warehouse developments can be expected and what affect these changes could have on internal control risks.

In addition to this, the study also provides reasons why the data warehouse environment can improve the efficiency of internal auditing when evaluating other routine cycles. This portion of the study focuses specifically on the data mining technique which allows the internal auditor to effectively interrogate small or large volumes of data when reviewing other audit cycles.

The chapter concludes by referring to the results of the empirical survey.

3. Future developments and the effects on internal control risk

The following areas have been identified as future developments which could significantly change existing data warehouse environments (DCI, 1998: 1-3). In addition to describing the development, the study also identifies what affect these changes could have on internal control risks:

3.1 Closed-loop business performance management

3.1.1 Definition

This development will result in the data warehouse providing management with information on a real-time basis. Rather than the user requesting data trends, information systems will smartly identify those data elements of defined interest and provide results timeously to the user without any prompting.

3.1.2 Internal control risks

- Although real-time data provision allows for timeous decision making, it can result in the user being inundated with unnecessary data. This will occur in instances where the criteria used in identifying reported trends are not established according to stringent management standards.
- If not properly managed and monitored, real-time data provision can result in the over utilisation of system resources.

3. 2 Increased access to data warehouse information

3.2.1 Definition

Increased access to the data warehouse by a larger percentage of the organisation's personnel is another expected development. In addition to personnel access, it is expected that the data warehouse will also be accessible by suppliers and customers in the future.

3.2.2 Internal control risks

- The risk of confidential information being disclosed to unauthorised users may be due to incomplete or inconsistent application of access rights procedures and policies.
- Increasing the number of users who have access rights to the data warehouse environment can negatively impact the overall performance of environment. This risk is increased if ongoing monitoring and the necessary upgrading of software and hardware is not conducted.

3.3 Removal of source data quality problems automatically

3.3.1 Definition

Data quality will improve as methods are developed to move from unclean data detection to the automatic cleansing of source quality problems up-front by the data warehouse interfaces and applications.

3.3.2 Internal control risk

The fact that any procedure which will be able to correct source data automatically within source applications without any form of stringent management control and audit trail is concerning. Such automatic procedures could result in unauthorised changes being made to source data.

3.4 Re-engineering the development methodology

3.4.1 Definition

Data warehouse requirements have been defined after the development of the associated source systems. It is also expected that the future source systems will be developed in such a manner that limited interfaces and data cleansing will be required.

This can be achieved by ensuring that data warehouse requirements are considered as part of the development methodology for traditional applications.

3.4.2 Internal control risk

There is a risk that the development process may become so inwardly focused that system development methodologies will not consider end user needs as the primary input in the development process.

3.5 Transferring of report and query functionality

3.5.1 Definition

Reporting and other query functions included in the data warehouse's source applications will be removed and will become functionality provided by the data warehouse environment.

3.5.2 Internal control risks

- The data warehouse will need to be accessed by a far larger audience. Therefore the risk of unauthorised access to privileged information is increased.
- User training and data warehouse package licenses are necessary and will increase costs.

4. Utilisation of data warehouse technology by the internal auditor

4.1 Background

There has been an increase in the use of data warehousing by internal auditors when evaluating internal controls within other audit cycles (Geiger, 1997: 31-32). In this section we will consider the use of data warehousing by the internal auditor, define the concept of data mining and finally provide a framework on how data warehousing can be used by the internal auditor in extracting needed audit evidence.

The major reasons why this technology has in recent years received so much more attention is based on the following developments (Moxon, 1996: 2):

- There has been a movement towards identifying information as the key corporate asset.
- Data mining provides a tool which can navigate through data and provide the exact level of detail without the need for technical assistance.
- The dramatic increase in hardware power and cost reductions therefore has provided the internal auditor with the ability to analyse and evaluate large volumes of transactions.

It may seem initially that the data warehouse cannot be relied upon by the internal auditor because the data has been manipulated. To ensure that the integrity of information is in its most objective form, the data should be obtained as close to the source system as possible (Geiger, 1997: 31).

There are however situations where data warehouse technology can be utilised effectively by the internal auditor in evaluating internal control risks within other routine audit cycles (Geiger, 1997: 31-32):

- Gathering data from multiple sources and then combining the data to obtain sensible evidence can be time-consuming and frustrating for the internal auditor. In the case of a data warehouse, the combination of all strategic information has already been performed and is a source which can be relied upon. The internal auditor should however ensure that he/she verifies the integrity of data relied upon. Steps which are performed to verify the data include evaluating the process adopted in migrating the data from its source to the data warehouse and identifying what steps management have implemented to ensure data integrity.
- The data warehouse environment includes metadata, therefore it can assist the internal auditor in simplifying interrogation and analysis tasks by providing information on how similar data from various outside sources compare.
- Data is quality checked before being permitted access to the data warehouse environment. Therefore information on data rejected provides the internal auditor

with a good indication of possible system and operational processes which require future auditing.

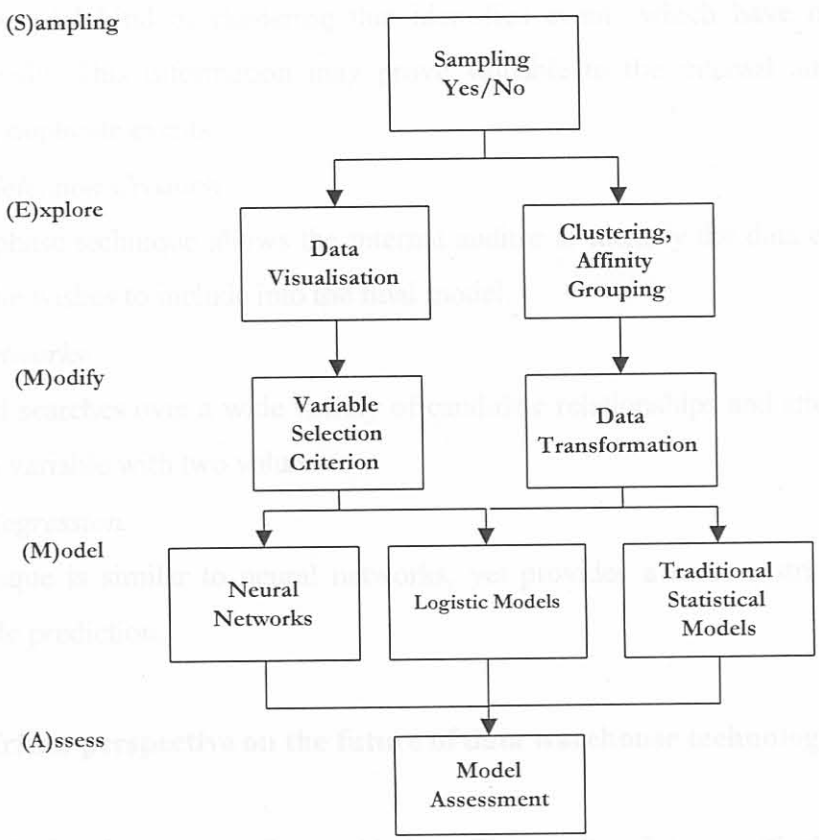
4.2 Defining data mining

Data mining is the automated analysis of customer transaction data for the purposes of discovering underling trends. Data mining as with all other components of the data warehouse environment is incorporated into the Knowledge Discovery in Databases process (Rosen, 1997: 51).

4.3 Process followed in utilising data warehouse technology

Data mining is effectively applied when a consistent methodology of evaluation and interrogation is adopted. To ensure the reliability of results and provide the internal

Figure 6.1 - SEMMA data mining process



Source: Casarin, 1997: 44

auditor with credible audit evidence, it is vital that a suitable methodology is adopted. The SEMMA methodology developed by the SAS Institute has proved to be the most frequently used of these models. The SEMMA acronym is based on the SAS's five stages in the data mining process, viz. Sampling, Exploration, Manipulation, Modeling and Assessment (Casarin: 1997: 44).

Figure 6.1 depicts the SEMMA process. Details on the various terms used in the SEMMA process include (Casarin, 1997: 44-46):

- *Data Visualisation*

  This involves representing data graphically to simply evaluations.

- *Clustering*

  This second phase function allows the user to examine large volumes of transactions and identify whether they can be grouped based on common criteria. This is also often termed "undirected data mining", because the user has no predetermined objective and is hoping that the data mining tool will reveal significant trends.

- *Affinity Grouping*

  This is a special kind of clustering that identifies events which have occurred simultaneously. This information may prove valuable to the internal auditor in identifying duplicate events.

- *Variable Selection Creation*

  This third phase technique allows the internal auditor to identify the data elements which he/she wishes to include into the final model.

- *Neural Networks*

  This model searches over a wide variety of candidate relationships and attempts to highlight a variable with two values.

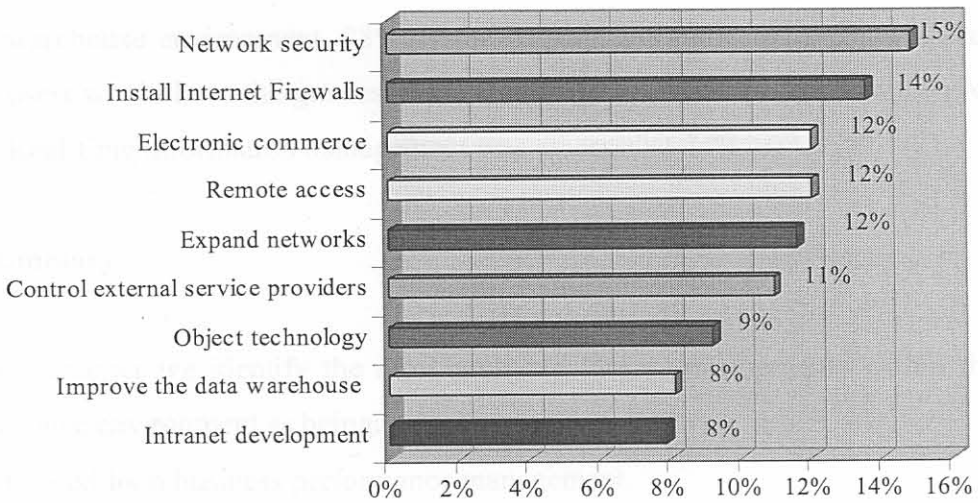
- *Logistic Regression*

  This technique is similar to neural networks, yet provides a more restricted and interpretable prediction.

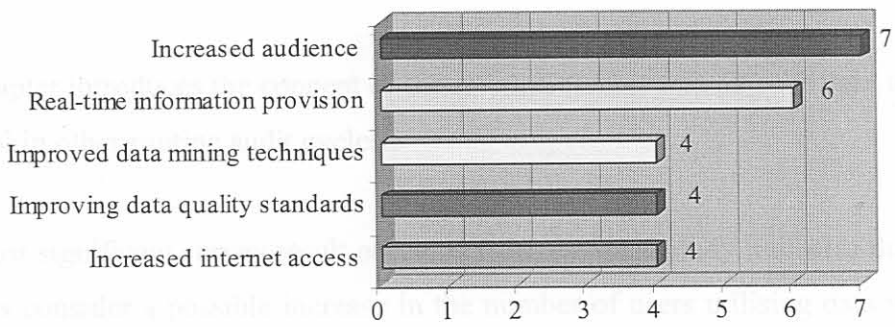
5. A South African perspective on the future of data warehouse technology

The results of the local survey are featured below. The results relate specifically to the future trends within data warehouse technology:

1. *Which of the following areas have been identified as the major risk applications and/or IT related systems within the next two years?*



2. *Which of the following data warehouse developments do you think will significantly affect your future assessment of the control environment?*



The most significant findings include:

- 15% of the respondents indicated that they expected network security to be the greatest area of concern within the next two years. Electronic commerce, remote access and expanding networks were rated at 12% individually. 8% of the

respondents felt that the data warehouse environment would be a significant area of concern in the foreseeable future. The most probable explanation for this low rating could be that management have not identified the data warehouse environment as a critical area as yet.

- Internal audit managers were requested to indicate which specific data warehouse elements they felt would significantly impact their assessment of the data warehouse environment. 28% of the respondents indicated that the increase in users would have the greatest impact on their assessment of internal control risk. Real-time information management was assessed at 24%.

6. Summary

In this chapter we identify the most probable future developments within the data warehouse environment as being:

- Closed-loop business performance management.
- Increased access.
- Improved data standards.
- Re-engineering the development methodology.
- Transferring of report and query functionality.

An explanation of the possible affect these developments could have on internal control risk was also provided.

The chapter introduces the concept of the internal auditor utilising the data warehouse as a tool in other routine audit cycles.

The most significant survey result obtained from the local study indicates that internal auditors consider a possible increase in the number of users utilising data warehouse technology as the factor which would most affect their assessment of internal control risk within the data warehouse environment.

7. Conclusion

The results obtained from the empirical survey indicate that the improvement of the data warehouse environment is considered to be the eighth most significant high risk area out of a possible nine selections. It is therefore imperative that internal audit teams ensure that they are equipped with suitable knowledge to evaluate the impact these future developments might have on the data warehouse environment. Only through close liaison with the management team and end user, will the internal auditor be able to effectively assess the changing data warehouse environment and provide value adding input to the organisation's operations.

Chapter 7

Conclusion

1. Summary

The purpose of this study is to identify the most pertinent internal control risks within the data warehouse environment. Fundamental concepts of the data warehouse have been introduced and a brief indication given of the key internal control risks within the overall data warehouse environment.

The study investigates the concept of a system development life cycle unique to the data warehouse environment. Reasons for the distinction between the system development life cycle for traditional systems and that for the data warehouse development are provided. Internal control risks specific to each of the phases within the data warehouse system development life cycle have been introduced. Suitable internal control considerations were provided as a means of assessing the extent of internal control risks. The most significant result of the empirical study indicates that organisations are utilising a development methodology unique to the data warehouse environment.

Six internal control risks which exist within the established data warehouse environment have been identified. Certain internal control risks identified include, not ensuring the completeness of data migrated to the data warehouse; inability to measure data quality; data warehouse access is not restricted to authorised users; etc. The most significant result of the empirical study indicates that internal auditors are relying on internet resources as a means of preparing for data warehouse reviews.

The concept of the dependent data mart environment has been introduced. The data mart is defined as an extension of the data warehouse. It allows for improved access to data by segregating data types according to business functions or subjects.

Two unique internal control risks have been identified. These are, a lack of sufficient response time monitoring on a periodic basis and the uncontrolled transfer of data from the organisation wide data warehouse to the dependant data mart. Suitable internal control considerations have been provided to assess the extent of these two unique internal control risks.

The study considers the concept of a distributed data warehouse environment. The internal auditor has been provided with a brief synopsis of the development and access and security considerations which should be applied during the development of a distributed data warehouse. Inability to restrict access, ensuing ongoing availability of the distributed data warehouse and the inefficient processing of queries within the distributed data warehouse, have been identified as the three fundamental internal control risks. Suitable internal control considerations have also been provided so that the internal auditor could assess the extent of the internal control risks identified.

Finally, the study concludes by highlighting the expected future trends and developments within the data warehouse environment. Attention has been drawn to defining each of the five areas where significant enhancements to the data warehouse are expected. The internal auditor has been provided with an indication of the internal control risks which could exist should the expected enhancements take place in the foreseeable future. The concept of the internal auditor utilising the data warehouse as a tool in other routine audit cycles is also introduced.

2. Further areas of research

In chapter 1, we said that little attention had been given to evaluating the data warehouse environment by the internal audit profession. The impact this evolving technology could have on the assessment of internal controls has yet to be explored.

The work presented in this study has only provided a brief insight into identifying and assessing the internal control risks within the various components of the data warehouse environment. Accordingly, two additional areas of research have been highlighted which would require more focused attention:

- *A comprehensive auditing framework is needed to assess internal control risks within decision support systems*
  By developing this framework, insight would be gained as to how the internal auditor could effectively assist management in ensuring that risk identification follows a consistent approach. All significant threats and possible opportunities would be identified and optimised.

- *The data warehouse environment needs to be practically integrated into the internal auditor's existing audit process as a means of identifying unexpected trends and irregularities when performing routine audits*
  Chapter 6 briefly introduces the concept of internal auditors utilising the data warehouse environment as part of their evaluation procedures when examining other routine audit cycles. It is evident from this brief study, that a more detailed investigation is needed to illustrate how this technology could be practically integrated into the traditional audit process. Any technology which would allow the internal auditor to perform his or her functions more efficiently and with increased rigour should be further investigated.

3. Conclusion

The following overall recommendations arise from this study:

- The data warehouse development team should ensure that a system development life cycle unique to the data warehouse environment is utilised whenever such an environment is being developed.
- Internal audit should be involved throughout the development process to ensure that significant internal control risks are identified.
- Internal auditors should ensure that management have considered the effect of future developments within the data warehouse environment when assessing the overall risk of such an environment.
- Internal auditors should use the data warehouse environment to identify significant trends and irregularities when performing routine audits.

In conclusion, the data warehouse environment provides unique opportunities for the organisation to ensure a more reliable and consistent decision making process. As the

reliance on systems such as the data warehouse environment increase, the need for a more controlled system is increased. Management and the internal auditor should work together to ensure that all significant internal control risks are identified timeously and that measures are implemented to negate the effects of these risks.

Annexure 1
Data Warehousing Survey Questionnaire

Name of person completing the questionnaire		
Designation		
Organisation's name		
Your office telephone number		
Organisation's staff size (*)	1-500	
	500-1500	
	1500-3000	
	3000-5000	
	5000-10000	
	> 10000	

(*) - Please mark with an "X" in the appropriate box alongside the correct value.

Completion instructions:

1. *Please note that no individual's name or title will be disclosed in the final results presented as part of the thesis submitted to the University of Pretoria, South Africa.*
2. *Please mark the appropriate answer with an X in the last column of the tables below except for the last question.*
3. *If your organisation does have a data warehouse/data mart structure in place, please complete sections A, B, C of the questionnaire. If however you do not have a data warehouse structure in place, but are planning on implementing such a structure, please complete Section B and C of the questionnaire.*
4. *Data warehouses and data marts are considered synonymous for the purposes of this questionnaire.*

Section A: An Already Implemented Data Warehouse Environment

Question	*Possible Answers*	*(X)*
1. For how long has your organisation had a data warehouse structure in operation?	Less than 1 year	
	1-2 years	
	2-5 years	
	5-10 years	
Comments?		
2. Which levels of staff within your organisation utilise the data warehouse structure (more than one reply is acceptable)?	Director level	
	Senior manager	
	Supervisory staff	
	Line personnel	
	Not sure	
Comments?		
3. Is the data warehouse environment identified as an application reviewed by internal audit on a periodic basis?	Yes	
	No	
Comments?		

Data Warehousing Survey Questionnaire
(continued)

Question	*Possible Answers*	*(X)*
4. If yes to question 3: How often will the data warehouse environment be audited by internal audit or appointed consultants?	Every year	
	Every 2 years	
	Every 3-5 years	
	Management request	
Comments?		
5. If yes to question 3: Which aspects of the data warehouse environment will be reviewed as part of the internal control review (definitions of the criteria used are included at the end of the questionnaire - more than one reply is acceptable)?	Confidentiality of information	
	Integrity of information	
	Reliability of Information	
	Availability of Information	
	Efficiency with which data is introduced into the data warehouse	
	Effectiveness of information in attaining management's initial goals	
	Compliance with data standards	
	Other (please include under comments)	
Comments?		
6. After completion of the post implementation review, did management realise the expected benefits of the data warehouse (if partially, express as a percentage of total under comments)?	Yes, fully	
	Partially	
	No	
Comments?		
7. Following on from question 6: If management did not realise/partially realise the expected benefits of the data warehouse, which of the following was considered the major cause for the failure?	Unrealistic expectations	
	Lack of sufficiently qualified personnel	
	Lack of funds	
	Unrecoverable/incomplete data from feeder systems	
	Other (please include under comments)	
Comments?		

Data Warehousing Survey Questionnaire
(continued)

Question	*Possible Answers*	*(X)*
8. Was the data warehouse implementation completed on time?	Yes	
	No	
Comments?		
9. If no to question 8: What was the major cause for the implementation not meeting the expected deadline?	Additional functionality requested by users after scope approval	
	Lack of qualified personnel	
	Lack of funds	
	Insufficient buy in from users	
	Unexpected system complexity	
	Other (please include under comments)	
Comments?		

Data Warehousing Survey Questionnaire
(continued)

Section B: The Data Warehouse Development

This section applies to both organisations that have completed their data warehouse development and those who are in the progress of developing such an environment.

Question	*Possible Answers*	*(X)*
1. Did the MIS/IT department develop a system development methodology specific for the data warehouse environment?	Yes	
	No	
Comments?		
2. On what basis were all possible "feeder" systems which could impact of the comprehensiveness of the data warehouse determined?	A source analysis prepared jointly by heads of department	
	Management knowledge	
	Data warehouse intended only for a single category of data	
	Other (please include under comments)	
Comments?		
3. What involvement will/does internal auditor play in the system development life cycle of the data warehouse?	Full project involvement	
	Integration testing only	
	Security aspects	
	No involvement	
	Other (please include under comments)	
Comments?		
4. What methodology was applied in ensuring that uniform data was introduced into the data warehouse environment?	A data quality standard was developed applied to each feeder system	
	Automated data extraction and conversion tools were utilised	
	Other (please include under comments)	
Comments?		
5. Was capacity management identified as an area which needed to be addressed as part of the development cycle?	Yes	
	No	
Comments?		

Data Warehousing Survey Questionnaire
(continued)

Question	*Possible Answers*	*(X)*
6. If yes to question 5: Which one of the following reasons was the major cause for including capacity management as part of the development cycle?	Scalability to handle future enhancements	
	Extensibility ensuring that the data warehouse can adapt to new hardware	
	Storage costs are monitored to verify return on investment	
	Other (please include under comments)	
Comments?		
7. What was management's major intention in implementing the data warehouse structure	Improved long-term focus	
	Empowering lower level management	
	Improved marketing strategy	
	Determining core competencies	
	Other (please include under comments)	
Comments?		
8. Will/has the data warehouse environment been included in the organisation's continuity plans/efforts?	Yes	
	No	
Comments?		
9. Will the data warehouse environment be distributed in nature?	Yes	
	No	
Comments?		

Section C: Audit Considerations

This section should be answered by all respondents.

Question	*Possible Answers*	*(X)*
1. What audit source materials were/will be used in formulating a suitable audit approach and program?	Data warehouse product manual	
	COBIT/COSO	
	Other ISACA/IIA source material	
	Internet sources	
	Other (please include under comments)	
Comments?		

Data Warehousing Survey Questionnaire
(continued)

Question	*Possible Answers*	*(X)*
2. Which of the following data warehouse developments do you consider will significantly affect your future assessment of the control environment (more than one reply is acceptable)?	Increased access via the internet and open communication channels	
	Real-time information provision	
	Increasing the audience of the data warehouse environment	
	Improving data quality standards	
	Improved data mining techniques for management use	
	Other (please include under comments)	
Comments?		
3. Which of the following areas have been identified as the major risk applications and/or IT related systems within the next two years (please use the scale 1 to 9. 1 being the most important and 9 being the least important).	Expanding network systems	
	Controls over external service providers	
	Deploying object technology	
	Implementing a data warehouse	
	Controls over electronic commerce	
	Deploying an intranet	
	Installing firewalls	
	Controls over remote access	
	Improve network security	
	Other (please include under comments)	
Comments?		

Data Warehousing Survey Questionnaire
(continued)

Criteria	*Definition*
Effectiveness	Deals with the information being relevant and pertinent to business process as well as being delivered in a timely, correct, consistent and usable manner.
Efficiency	Concerns the provision of information through optimal (most productive and economical) use of resources.
Confidentiality	Concerns the protection of sensitive information from unauthorised disclosure.
Integrity	Relates to the accuracy and completeness of information as well as to its validity in accordance with business values and expectations
Availability	Relates to information being available when required by the business process now and in the future. It also concerns the safeguarding of necessary resources and associated capabilities
Compliance	Deals with complying with those laws, regulations and internal quality standards for data.
Reliability of information	Relates to the provision of appropriate information for management to operate the entity and for management to exercise its financial and compliance reporting responsibilities.

Source: ISACA:1998.

Any other comments:

_____

_____

_____

_____

Would you like an e-mail copy of the thesis once completed? _____

If yes, please provide your e-mail address: _____

Thank you for taking the time in completing the survey. Please contact me if you require any other information.

Contact Person: *Sean de la Rosa*
Phone: *+27 11 490 0675*
Fax: *+27 11 493 1580*
Mobile: *+27 82 82 13471*
e-mail *sean.delarosa@afrox.boc.com*

Annexure 2

Strategic Data Warehouse Checklist

Criteria	*Guidelines*	*Response*
i. What are the number of internal source systems and databases?	Each source system along with their databases and files will take additional research, including meetings with those who have knowledge of the data. The time to document the results of the research and meetings should also be included in the time estimates.	
ii. How many business processes are expected for the data warehouse project (Examples: analyse sales, markets and financial accounts).	A project should be limited to just one business process. If management insist on more than one, the time and effort will be proportionally higher.	
iii. How many subject areas are expected for the project (Examples: customer, supplier / vendor, store/ location).	If possible, a project should be limited to just one subject area. If management insist on more than one, the time and effort will be proportionally higher.	
iv. Will a high level enterprise model identifying all possible systems be developed during the project?	Ideally, an enterprise model should have been developed prior to the start of the data warehouse project. If the model has not been finished and the project requires its completion, the time scheduled must be adjusted.	

Strategic Data Warehouse Checklist (continued)

Criteria	*Guidelines*	*Response*
v. How many attributes (i.e. fields and columns) will be selected for the project?	The more attributes to research, understand, clean, integrate and document, the more resources that will be needed.	
vi. Are the source files well modeled and documented?	Documentation is critical to the success of the project. Extra time and effort must be provided for if the source files and databases have not been well documented.	
vii. Will there be any external data in the project, and if so, is it well documented?	External data is often not well documented and usually does not comply with organisation standards. Integrating data is often difficult and time consuming.	
viii. Is the external data modeled (i.e. accurate, actively being used and comprehensive, etc.)?	Without a model, the effort to understand the source external data is significantly greater. It is often unlikely that the external data has been modeled.	
ix. How much cleaning will the source data require?	Data cleaning both with and without software tools to aid the process is tedious and time consuming. Organisations usually overestimate the quality of their data and may underestimate the effort to clean the data.	

Strategic Data Warehouse Checklist (continued)

Criteria	*Guidelines*	*Response*
x. How much integration will be required?	The need to integrate various source systems from various data stores can require significant resources (especially when complex external data elements must be introduced into the data warehouse environment).	
xi. What is the estimated size of the data warehouse database?	Databases in excess of 500 gigabytes may generate performance problems. These concerns should be kept in mind when considering costs and service level requirements.	
xii. What are the service level requirements?	In instances where the data warehouse will be required five days a week, eight hours a day, etc. significant running and maintenance costs will be incurred.	
xiii. How frequently will data elements need to be loaded and updated?	In instances where uploads are required more frequently, an increased effect on performance will be noted.	
xiv. Will a new hardware platform, user PC's or network infrastructure be required? If so, will it be different than the existing hardware?	The installation of new hardware will require planning, operations training and familiarisation.	

Strategic Data Warehouse Checklist (continued)

Criteria	*Guidelines*	*Response*
xv. How many query tools will be chosen?	The acquisition of new query tools also involves significant training and needed support time.	
xvi. Is user management sold on and committed to the data warehouse project and what is the organisational level at which the commitment was made?	If management are not sold on the project, the risk is significantly greater. Risks could include increased difficulty in getting resources and timely responses from affected parties.	
xvii. Who do the data warehouse project manager's report to?	The higher up the project manager's report, the greater the commitment and the more visibility the project will receive.	
xviii. Will the appropriate users be committed and available for the data warehouse project?	The more users which will be needed to provide input into the project, the more difficult coordination efforts will be. Also, if people important to the project are not committed and available, it is unlikely that the project will be completed on time.	
xix. Will knowledge application developers be available for the migration process and system testing?	Consideration should be given to ensuring that a sufficient number of application developers can be obtained to assist in the data warehouse development.	

Strategic Data Warehouse Checklist (continued)

Criteria	*Guidelines*	*Response*
xx. Will the database administrators be familiar with the chosen database management system and be available for the project duration?	As with all new software, a new database management system will require recruitment. It is also vital that the database administrator is available for the full duration of the data warehouse project.	
xxi. Will technical support people be available for capacity planning, performance monitoring and trouble shooting?	In instances where the organisation is considering an organisational-wide data warehouse, the need for support people will increase. Support in issues such as capacity planning and performance monitoring will become more critical as the data warehouse increases in size and complexity.	
xxii. How many queries are expected and what is the anticipated level of complexity of such queries?	The higher the volume of queries and the greater their complexity, the more user training that will be required.	
xxiii. Are there any significant security and/or audit issues which must be considered before the data warehouse is implemented?	Audit involvement in the data warehouse project is recommended and will require audit's commitment.	

Source: Adapted from Adelman, 1998: 1-4

Annexure 3

Roles and Responsibilities of the Data Administrator

The roles and responsibilities of the data administrator are sub-divided into four areas of specialisation. These areas of specialisation are detailed below as well as the associated objective of each. The roles and responsibilities are also included under each of the functional areas identified.

1. Data Administration Infrastructure

A. Objective of function

To provide the support framework for definitions, use and maintenance of data resources.

B. Roles and responsibilities

i. Establish and maintain the appropriate data administrator infrastructure for each customer type

This includes establishment of data administrator organisation, roles, responsibilities, external liaison interfaces and data administrator/repository administration internal procedures. A mission statement or charter needs to be written to align the four functions and new services with customer data needs. Job descriptions must reflect the role(s) staff members have in the new four functions and the skill sets required for such roles.

ii. Coordinate centralised/decentralised data administrator functions

Revised data administrator functions may remain centralised at the enterprise level, yet decentralised at the application project level to meet customer needs. Application data administrators perform data modeling roles and limited project repository administration roles to leverage data administrator activities within the development project. Decentralised application data administrator roles are to be coordinated by the centralised data administrator function.

iii. *Establish and maintain data administrator tools and tool roles*

Evaluation and recommendation of tools are conducted with other appropriate organisations. Once the tool has been determined, the appropriate infrastructure surrounding that tool must be established (i.e., repository/tools architecture, training, tool roles, tool security, etc.). The appropriate data/metadata policies, standards, procedures for each customer type needs to be enhanced and/or developed.

iv. *Plan strategically for future data/information needs*

This involves coordinating with other enterprise planning efforts to develop and publish a strategic long-term (usually five years ahead) data/information plan. This long-term plan provides target strategies, policies, standards, models and software tools. A migration plan of steps to take for reaching the target goals should be included. This strategic information plan feeds the data administrator organisational short-term plan.

v. *Establish and maintain a data administrator organisational short-term plan*

The initial data administrator short-term plan (two years or less) coordinates customer types, upcoming projects and data administrator roles.

vi. *Communicate data administrator topics*

Communicate data administrator topics to each customer type by giving presentations, training, publishing on the Web, etc. Data and metadata concepts as well as the appropriate policies, standards and procedures need to be disseminated to customer groups.

vii. *Establish and maintain standard practices*

Establish and maintain standard practices with regard to data analysis approach, data element identification, deliverables and tools and techniques for all customer types. Standard practices include the policies, standards and procedures that are enforced to minimise redundancy. It also includes data standardisation and data quality techniques that analyse mapping of data elements and constructs across databases.

2. Data Model Administration

A. Objective of function

To support creation and maintenance of the data architecture.

B. Roles and responsibilities

i. Create, publish and maintain the data architecture

This includes subject areas identification as well as creation and maintenance of the enterprise or corporate data model that contains high-level entities and their relationships derived from business rules.

ii. Project support

Provide project support for each customer type through the development and/or the review of data models.

3. Repository Administration

A. Objective of function

To develop and maintain the repository environment that each customer type should be encouraged to use. Data administrators should work toward the goal of having the repository become the centralised location of metadata for both the development and production environments as well as for end-user data access.

B. Roles and responsibilities

i. Establish and maintain the repository model architecture and repository objects (meta model)

The repository/tools architecture is a diagrammatic view of all tools and applications that fall within the scope of the initial repository interface requirements. It provides an essential road map of the way the repository environment should operate in a multitiered environment. Also, every

repository comes with a meta model with related entity types, policies, methods, templates, security, migration, etc. The data administrator should establish a meta model customised to its customer environment for use in the repository.

ii. *Enforce naming standards, keywords, abbreviations*

Manual enforcement of established repository standards is difficult and time-consuming. What is preferable is automated tool enforcement through proper implementation of edits that represent the standards. This step of validating names, keywords, abbreviations, etc., must be built into the systems development life cycle as part of the data administrator review activities.

iii. *Establish and maintain ongoing repository administration for each data administrator customer type*

This involves communicating and advising users on what exists in the repository and how to access it. Repository controls and security profiles must be established to protect, access and change repository components. Templates and standard reports need to be developed for customers to view repository contents and obtain reports.

4. End-User Data Access Administration

A. *Objective of function*

To create and maintain the data access categories and support the data warehouse and other data access projects.

B. *Roles and responsibilities*

i. *Create and maintain subject areas and relationships*

Create and maintain subject areas and relationships to entities for each customer by producing data for end-user access. Data warehouse or data mart projects are designed for end-user data retrieval and analysis. The enterprise may provide external data files for public use. Each data access project needs a

meta access model for the selected data access browser. This model contains the identified major and minor categories of entities for ease of use and data retrieval.

ii. *Provide data warehouse support*

Provide data warehouse support in the areas of modeling, source data quality, business rules, end-user tools, etc. Models need to be created, reviewed and maintained for each data warehouse. Data administrator support can include identification of source data to feed the warehouse as well as creation and maintenance of source-to-target data mappings, transitional data source-to-target rules and data access rules. The data administrator should establish the appropriate customer data warehouse/data access metadata directories that interface and integrate with the established repository environment.

Source: Adapted from Cupoli, 1999: 1-5

Annexure 4

Vendor Prescreening and Application Selection Questionnaire

Criteria	*Guidelines*	*Response*
A. Vendor Information Business Profile		
• Where does the vendor see their company in the future? • What is the vendor's guiding philosophy? • Which sector of the industry does the vendor fall under (examples: data warehousing, middleware, etc.)? • Which industry need is addressed by the product (examples: extraction process, testing and verification of data quality, etc.)?	Vendor information will help the organisation predict the vendor's long-term competitive strength in the market. The vendor's vision, or lack thereof, is a reflection of their view of the future and will provide insight into the viability of a long-term relationship.	
Indicate the number of years that the vendor has been in business by providing the aforementioned product.		
How many clients does the vendor have where the evaluated product is being evaluated and/or is already installed?	The completed questionnaire should be returned with full details of 5 (or more) clients and their references with contact numbers. Reference sites should refer to the evaluated product.	

Vendor Prescreening and Application Selection Questionnaire
(continued)

Criteria	*Guidelines*	*Response*
B. Financial Profile		
• What has been the annual increase in turnover over the past 5 years? • Is the vendor organisation funded by personal capital or a publicly traded company? • Provide a recent copy of the vendor's financial statements (if allowable).		
C. Product Information Version Information		
• What is the current version number of the product? • When last was a new release of the product being evaluated, issued? • How often does the company come out with new releases of the product being evaluated?		
D. Environment Supported		
• Specify the hardware platforms on which the product will function as per the specification. • Specify the software products and their versions that the product has been tested for.		

Vendor Prescreening and Application Selection Questionnaire
(continued)

Criteria	*Guidelines*	*Response*
• Specify the network protocol and any other information regarding the network compatibility of the product (Examples: TCP/IP, Remote Procedure Call, etc.).		
For data warehouses: • Is the evaluated product metadata aware? • Can the evaluated product read and use external metadata? • Does the product provide an interface to the metadata interchange? • Does the product use metadata internally?	Metadata is retained within the data base management system and includes information about the structure, content, keys, and indexes of the migrated data.	
D. Dependency on Other Vendor Products		
List the products that are critical for functioning of the product or on which the evaluated product must be used.	Ensure that the vendor's product keeps up with the changing versions of software it may be dependent on. This is done to ensure that future application upgrades in a multiple-product dependent environment may not be adversely affected.	

Vendor Prescreening and Application Selection Questionnaire
(continued)

Criteria	*Guidelines*	*Response*
E. Quality Information		
• Describe the testing environment existing in the vendor's company for development and support of the product.	Vendors that subscribe to software process models such as CMM, ISO or Cleanroom provide an indication of their commitment to quality.	
• Briefly list the average defect rate over the life of the product.		
F. Product Pricing Information		
• Obtain a copy of the vendor's terms of conditions. • Identify the pricing per copy of software purchased. • Detail any licensing fees in addition to the previous costs. • Detail the annual support and maintenance fee charged. • Obtain details of the product's warranty period and terms. • Obtain details of any other miscellaneous charges (examples: call-out costs, transportation and delivery, etc.).		

Vendor Prescreening and Application Selection Questionnaire
(continued)

Criteria	*Guidelines*	*Response*
G. Technical Support and Service Commitment		
• Identify the different technical support schemes on offer by the vendor. • Define what services will be included in the service commitment agreements offered by the vendor.	The service commitment agreement should bind the vendor to provide a guaranteed level of service within a specified period of time. Failure to comply with sanctioned terms should result in some form of penalty being incurred.	
H. Consulting/Mentoring Training and Consulting		
Provide full details of the training and consulting services provide by the vendor and what the associated costs are.		
I. Year 2000 Compliancy		
• Verify that the product is Year 2000 compliant. • Request third party verifications and authentic test results verifying Year 2000 compliance.		
J. Scalability		
• Can the product allow additional clients and servers to be added and removed in the future to handle increases and decreases in load?		

Vendor Prescreening and Application Selection Questionnaire
(continued)

Criteria	*Guidelines*	*Response*
• Provide independent results on the product performance when increasing and decreasing data loads.		

Source: Adapted from Tiwary S., Tewary A., 1998: 1-5

Glossary

Access	The operation of seeking, reading, or writing data on a storage unit.
Algorithm	A set of statements organised to solve a problem in a finite number of steps.
Application	A group of algorithms and data interlinked to support an organisational requirement.
Archival Data	A collection of data of a historical nature. As a rule, archival data cannot be updated. Each unit of archival data is relevant to a moment in time.
Attribute	A property that can assume values for entities or relationships. Entities can be assigned several attributes. Some systems also allow relationships to have attributes as well.
Audit/Log trails	Data that is available to trace activity, usually update activity.
Authentication	The act of verifying the identity of a user and the user's eligibility to access computerised information (usually by means of password verification). Designed to protect against fraudulent logon activity.

Availability	Component of internal control risk. Availability relates to information being available when required by the business process now and in the future. It also concerns the safeguarding of necessary resources and associated capabilities.
Breadbox Analysis	A phase of the data warehouse development. An analysis which determines the volume of data which will be retained within the data warehouse environment.
Capacity Planning	An assessment used to determine whether the data warehouse and associated interfaces will be able to manage an increase in the volume of data transferred to the data warehouse.
Central Processing Unit	Computer hardware which houses the electronic circuits that control/direct all operations in the computer system.
Closed-loop Business Performance Management	Future functionality which may provide data warehouse information on a real-time basis.
Compliance	Component of internal control risk. Compliance deals with complying with those laws regulations and contractual arrangements to which the business process is subject, i.e. externally imposed business criteria.
Confidentiality	Component of internal control risk. Confidentiality concerns the protection of sensitive information from unauthorised disclosure.

Continuity Plan	A formalised procedure developed by management to assist in addressing loss of operations due to controllable and/or uncontrollable disasters.
Conversion Plan	The conversion plan details the mapping of data from the operational environment to the data warehouse environment. It also identifies the best route to migrate source data to the data warehouse.
Data Administrator	The individual responsible for the specification, acquisition, and maintenance of data management software and the design, validation, and security of files or databases. The data model and the data dictionary are classically the charge of the data administrator.
Database	A collection of interrelated data stored (often with controlled, limited redundancy) according to a schema. A database can serve single or multiple applications.
Database Administrator	The organisational function charged with the day-to-day monitoring and care of the databases. The database administrator is more closely associated with physical database design than the data administrator is.
Database Management System	A computer-based software system used to establish and manage data.
Data Element	(1) An attribute of an entity; (2) A uniquely named and well-defined category of data that consist of data

items and that is included in a record of an activity.

Dependant Data Mart — A database, or collection of databases, designed to help managers make strategic decisions about their business. Whereas a data warehouse combines databases across an entire organisation, data marts are usually smaller and focus on a particular subject or department.

Data Mining — The automated analysis of detailed operational customer transaction data for the purposes of discovering hidden, unidentified or underlying patterns.

Data Model — (1) The logical data structures, including operations and constraints provided by a database management system for effective database processing; (2) The system used for the representation of data.

Data Model Structure — A logical relationship among data elements that is designed to support specific data manipulation functions.

Data Warehouse — A collection of integrated, subject-orientated databases designed to support the decision support function, where each unit of data is relevant to some moment in time. The data warehouse contains the lowest level data and lightly summarised data.

Decision Support Systems — A system used to support managerial decisions. Usually they involve the analysis of many units of data in a heuristic fashion. As a rule, their processing does not involve the update of data.

Distributed Data Warehouse	A data warehouse environment dispersed over a large geographical area to address the unique needs of such a structured organisation.
Effectiveness	Component of internal control risk. Effectiveness deals with information being relevant and pertinent to the business process as well as being delivered in a timely, correct, consistent and usable manner.
Efficiency	Component of internal control risk. Efficiency concerns the provision of information through the optimal (most productive and economical) use of resources.
Encryption	A technique used to protect the plain text by coding the data with suitable algorithms such that it is unintelligible to the reader.
Extract	The process of selecting data from one environment and transporting it to another environment.
Firewall	Software used to protect against denial of services and any unauthorised access to internet resources. The system should control any application and infrastructure management flows in both communication directions, i.e. data sent and received by the organisation.
Granularity	The level of detail contained in a unit of data. The more detail there is, the lower the level of granularity. The less detail there is, the higher the level of granularity.

Heuristic	The mode of analysis in which the next step is determined by the results of the current step of analysis.
Index	The portion of the storage structure maintained to provide efficient access to a record when its index key item is known.
Information	Data human beings assimilate and evaluate to solve problems or make decisions.
Information Technology Department/Personnel	Personnel and/or department appointed as custodians of computer resources.
Interface	Program logic responsible for transferring data from the operational environment to the data warehouse environment.
Integrity	Component of internal control risk. Integrity relates to the accuracy and completeness of information as well as to the validity in accordance with business values and expectations.
Internal Control Objective	The specific goal of an audit or review. These often center around substantiating the existence of internal controls to minimise internal control risk.
Internal Control Risk	The risk that management's plans, organisation and associated procedures will not provide reasonable assurance that the organisation's goals and objectives will be achieved. See **Effectiveness, Efficiency, Confidentiality, Integrity, Availability,**

	Compliance and **Reliability** regarding components of internal control risk.
Key	A data item or combination of data items used to identify or locate a record instance (or other similar data groupings).
Key, Primary	A unique attribute used to identify a single record in a database.
Key, Secondary	A non-unique attribute used to identify a class of Records in a database.
Knowledge Discovery in Databases	A framework applied in decision support systems. The Knowledge Discovery in Databases process is split into five stages, viz. Selection, pre-processing, transformation, information access layer systems and, interpretation and evaluation.
Load	To insert data values into a database that was previously empty.
Metadata	(1) Data about data; (2) The description of the structure, content, keys, indexes, etc., of the data.
Migration	The process by which frequently used items of data are moved to more readily accessible areas of storage and infrequently used items of data are moved to less readily accessible areas of storage.
Network	A system of interconnected computers and the communications equipment used to connect them.

Operational Data — Data used to support the daily processing a company does.

Populate — To place occurrences of data values in a previously empty database. See **Load**.

Primitive Data — Data elements whose existence depends on a single occurrence of a major subject area of the enterprise. Primitive data is distinguished from secondary data thereby ensuring that duplicate data elements are avoided and that the most accurate data element is selected if duplicates are found. See **Secondary Data**.

Project Team — The data warehouse project team consists of a project leader, business analysts, data administrators, database administrators, systems support, computer programmers and end users. These personnel will be responsible for the overall project administration, design of the warehouse structures; analyse source data; identify how data is to be linked and, and if applicable, integrate external sources.

Quality Assurance Review — The aims of the quality assurance review are to determine whether the data warehouse development is being administered according to established standards and that project deficiencies are identified timeously and corrected with minimal resources.

Query Optimiser — A sub-module of the query processor used to govern and expedite the processing and data transmission required for responding to queries. Its aim is to ensure that either the total cost or the total response

	time for a query is minimised.
Record	An aggregation of values of data organised by their relation to a common key.
Redundancy	The practice of storing data more than one occurrence of data. In the case where data can be updated, redundancy poses serious problems. In the case where data is not updated, redundancy is often a valuable and necessary design technique.
Referential Integrity	The facility of a database management system to ensure the validity of predefined relationships.
Reliability	Component of internal control risk. Reliability of information relates to the provision of appropriate information for management to operate the entity and for management to exercise its financial and compliance reporting responsibilities.
Repository	See **Database**.
Return on Investment	A tool used for weighting expected benefits against the costs of a specific project. Return on Investment calculations include cash flow analysis, net present value calculations, return on investment and payback calculations.
Security Plan	A plan which identifies the organisation's standards regarding system security and the user's responsibilities while utilising computer equipment and/or system software.

Secondary Data	Data elements whose existence depends on two or more occurrences of a major subject. See **Primitive Data**.
SEMMA Methodology	A data mining methodology. It consists of 5 stages, viz. sampling, exploring, modifying, modeling and assessing.
Snapshot	A database dump or the archiving of data out of a database as of some moment in time.
Source Systems Analysis	A phase of the data warehouse development. An analysis which defines the source data elements, evaluates the accuracy of data before migration and considers how to manage the volume of data elements.
Stability Analysis	An analysis which considers the grouping together of data elements which will change based on similar conditions.
Strategic Planning	A phase of the data warehouse development. Strategic planning identifies whether the organisation is in need of a data warehouse environment, and if so, what would the extent of such a development be.
Subject Area Analysis	A phase of the data warehouse development. Identification of suitable population data from existing applications which should be introduced into the data warehouse environment.
System Development Life Cycle	The framework applied in the development of an application. The framework is split into distinct phases to improve controllability of the development

process.

Table	A relation that consists of a set of columns with a heading and a set of rows.
Technical Environmental Preparation	A phase of the data warehouse development. The primary aim of this phase will be to determine how the organisation's network capabilities will be affected by the increased traffic created by the data warehouse.
Time Stamping	The practice of tagging each record with some moment in time, usually when the record was created or when the record was passed from one environment to another.
Technical Assessment	A phase of the data warehouse development. It focuses on determining the architectural configuration needed to house the data warehouse.
User	A person or process issuing commands and messages to the information system.
Year 2000 compliance	Identified as the risk that hardware and/or software will be unable to process transactions with 4 digit year date fields (i.e. DD/MM/**YYYY**).