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.

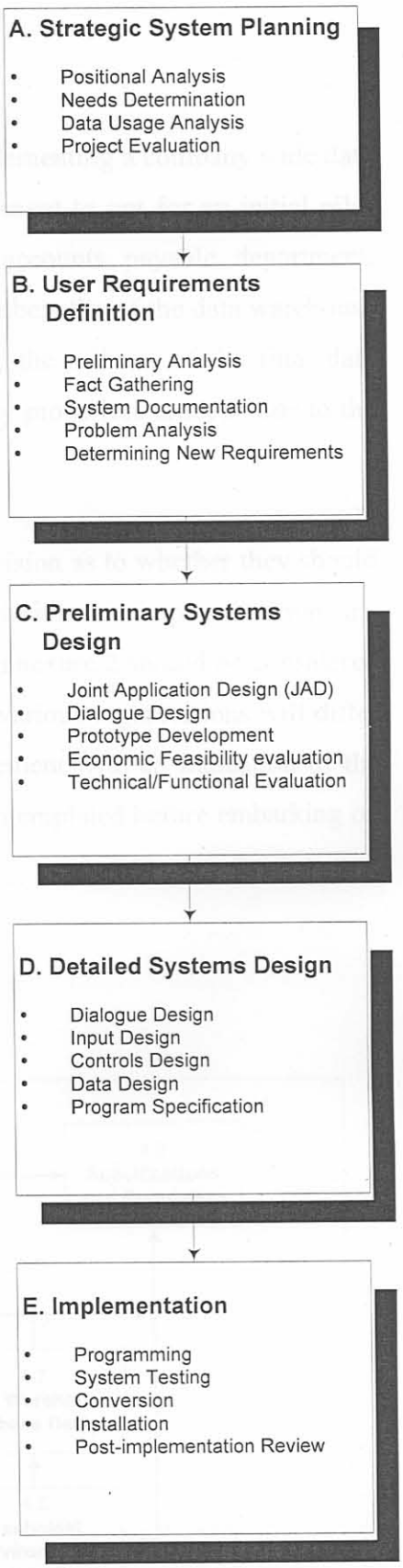
3. Why a different system development life cycle exists for data warehouses

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

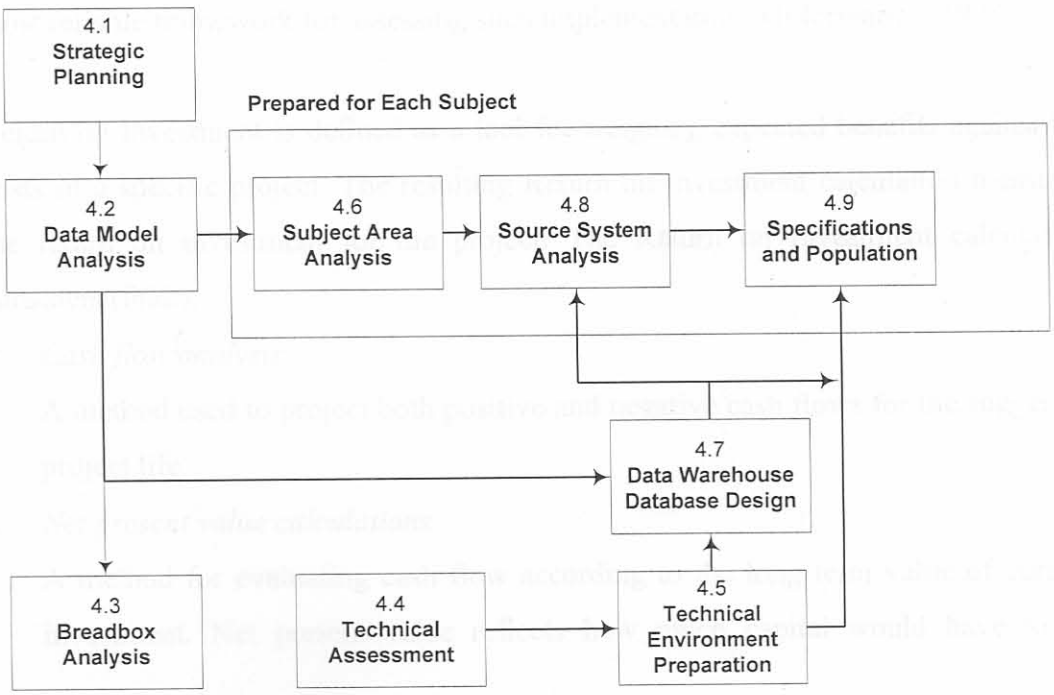
4.1 Strategic planning

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):

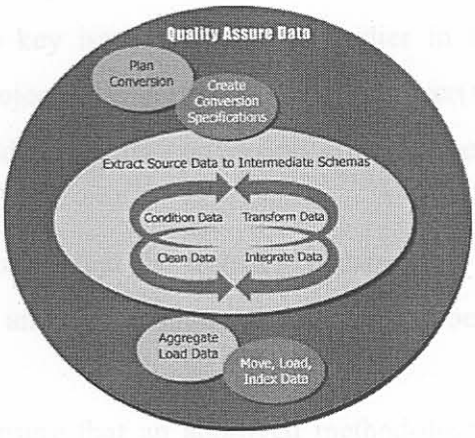
- *Definition of source data elements*
  This step attempts to verify that data labeled within the operational system retains as much of its original character as possible as origin must be easily traced once included into the data warehouse. This improves traceability and follow-up of inconsistencies 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:

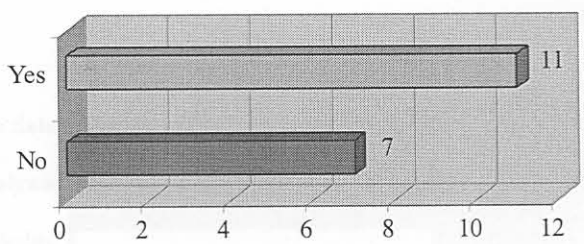
- The project team, in conjunction with the procurement section, should choose the most reputable vendor based on predetermined assessment criteria.

- 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.

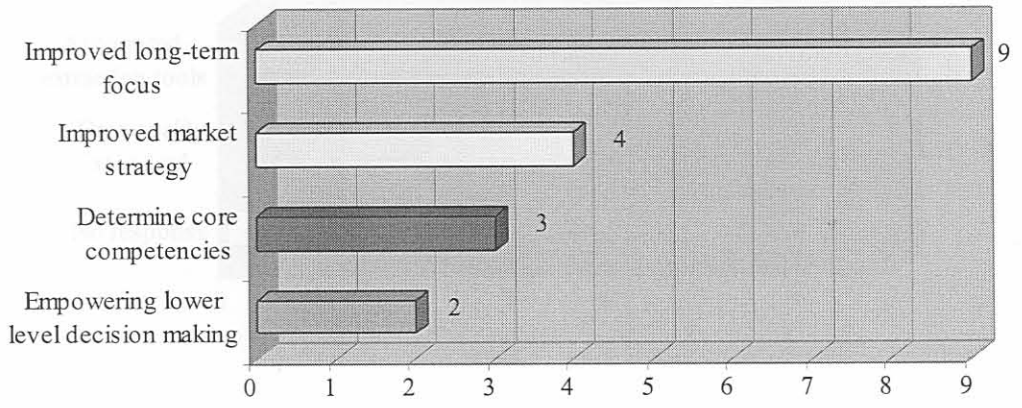
6. A South African perspective on the audit of developing data warehouse environments

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:

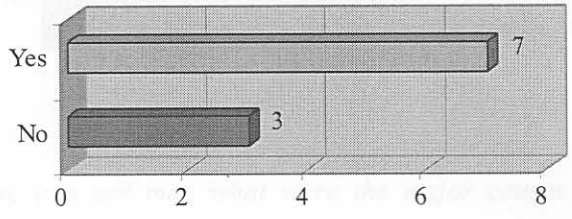
1. Does the company already have or is planning on implementing a data warehouse environment?



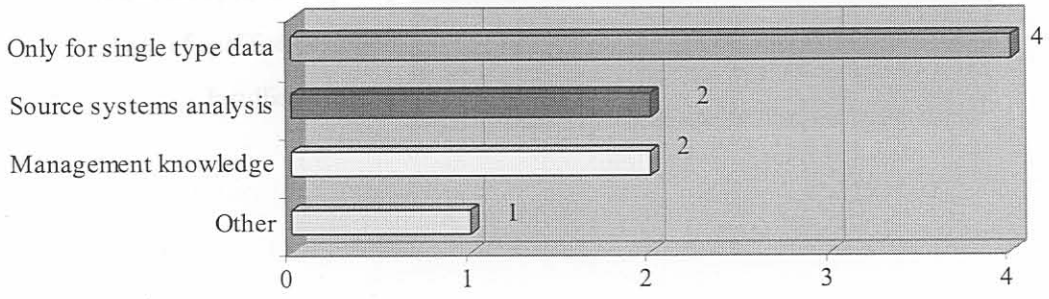
2. What was management's major intention in implementing the data warehouse?



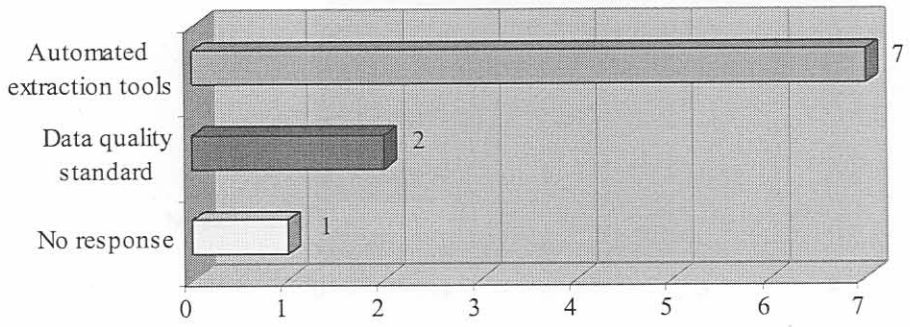
3. Did the Information Technology Department develop a system methodology specific for the data warehouse environment?



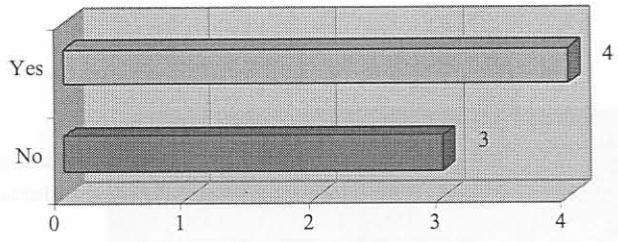
4. On what basis were all possible source systems identified?



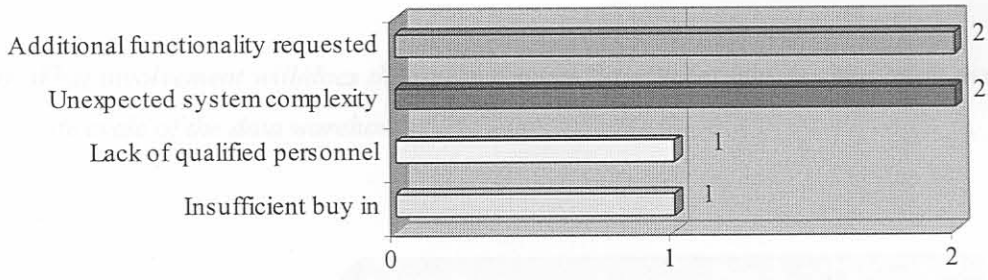
5. *What methodology was applied in ensuring that uniform data was introduced into the data warehouse?*



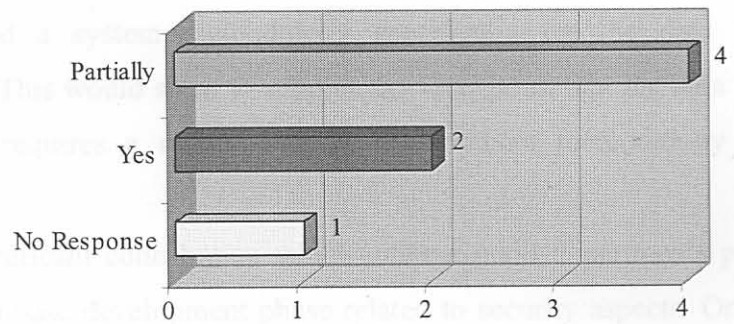
6. *Was the data warehouse implementation completed on time?*



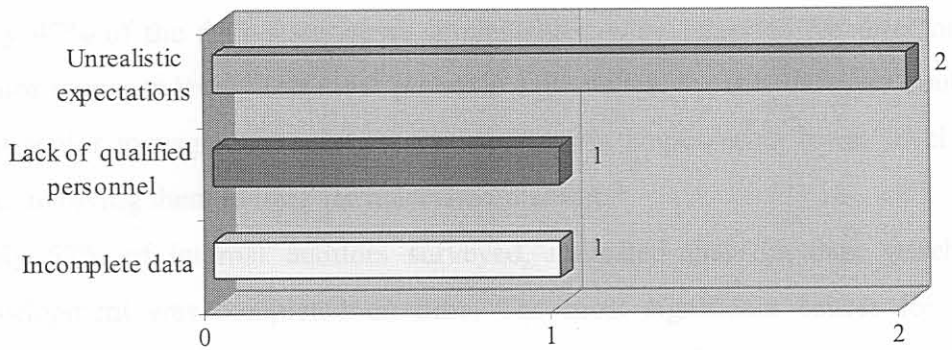
7. *If the deadline was not met, what were the major causes for the implementation not meeting the expected deadline?*



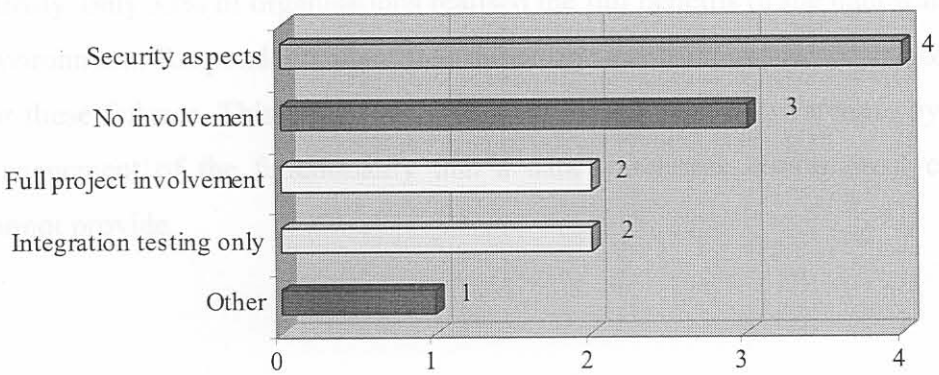
8. *Subsequent to the post implementation review, did management realise the expected benefits?*



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.