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

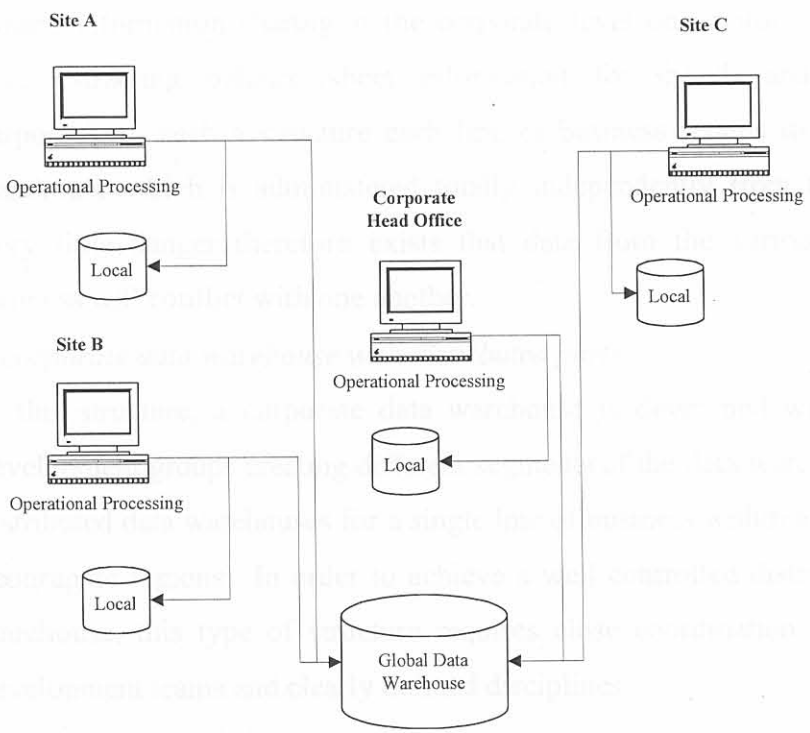
considered during an evaluation of such an environment. This chapter concludes by providing suitable internal control considerations which can be used in assessing internal control risk. The result of the empirical survey conducted as part of the study is also included.

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

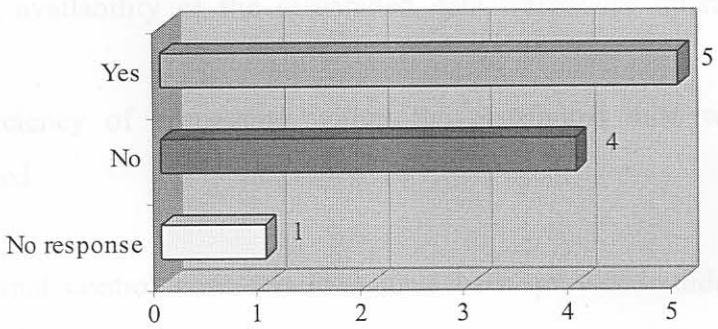
The following internal control considerations are applicable (ibid.):

- 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.