TOWARDS SECURITY IN DATA INTEGRATION: A FRAMEWORK

Wayne Routly[1] and Dalenca Pottas[2]

[1]Department of Computer Services, Port Elizabeth Technikon, South Africa
[2]Department of Applied Informatics, Port Elizabeth Technikon, South Africa

[1]waynear@petech.ac.za, +27 041 5043159, Private Bag X6011, Port Elizabeth, 6000
[2]dalenca@petech.ac.za, +27 041 5049100, Private Bag X6011, Port Elizabeth, 6000

ABSTRACT

The new millennium has brought about an increase in the use of business intelligence and knowledge management systems. The very foundations of these systems are the multitude of source databases that store the data. The ability to derive information from these databases is brought about by means of data integration. With the current emphasis on security in all walks of information and communication technology, a renewed interest must be placed in the systems that provide us with information; data integration systems.

In the database environment we are concerned with the database itself and the media used to connect to and from the database. In distributed data integration, the concept of the database is redefined to the source database, from which we extract data and the storage database in which the integrated data is stored. This postulates three distinct areas in which to apply security, the data source, the network medium and the data store. All of these areas encompass data integration and must be considered holistically when implementing security. Data integration is never only one server or one database; it is various geographically dispersed components working together towards a common goal. It is important then that we consider all aspects involved when attempting to provide security for data integration.

This paper investigates security issues in the data integration cycle, with special reference to problems when performing data integration in a peer-to-peer environment, as in distributed data integration.

KEYWORDS

Data Integration, Security, Data Integrity

TOWARDS SECURITY IN DATA INTEGRATION: A FRAMEWORK

1 BEFORE THE STORM

"We are in the business of selling databases, and you have bought too many of them, says Larry Ellison, CEO of Oracle. "Having lots of little databases all over the organization is extremely costly to maintain. It's very difficult to know where to look for the information when you need to make decisions." (Eastwood, 2001)

This statement, from one of the worlds leading database vendors is unsettling, to say the least. Throughout the 1990's companies have been expanding their knowledge bases through mergers, acquisitions and an increased focus on Customer Relationship Management (CRM) and Supply Chain Management (SCM). This has resulted in an information store for almost every business function. The enterprise landscape is littered with freestanding data stores and the myriads of information sources each new application creates. From an integrity standpoint this proves problematic as the enforcement of constraints that affect the consistency of the database is difficult in a distributed environment (Ibrahimm, 2002).

We have also seen an increase in the volume of data being sent across networks, this includes confidential information. This increase has gone hand in hand with an increase in the cases of data being compromised. A compromise to an organization's data is in actual fact a compromise to the business itself.

Data is the driving force behind the organizations of today. Loss of said data can have effects such as but not confined to:

- Loss of revenue for the company.
- A decrease in productivity due to data inconsistencies.
- Loss of customers due to confidentiality breaches.
- A decrease in customer loyalty due to data quality concerns.
- Possible bankruptcy due to breach of security

If organizations want to prevent these scenarios, it is paramount that both the security and the integrity of their data become a priority.

2 ENTER DATA INTEGRATION

Organizations are becoming aware of the importance of the information assets they posses. They need to be able to respond to customer needs and market conditions within a very brief window of opportunity. Historically, it was common to manage each data source as though it were a self-contained world. However, this practice is no longer in line with the business functions and how decisions are made.

The solution that organizations are using to unify the maze of information wealth they posses, is data integration. Data integration allows for a 'single view' of the organization's information. It is the problem of combining the data residing at different sources, and providing the user with a unified view of the data, called a global or mediated schema (Halevy, 2001). It is a way of providing

uniform access to a series of collections of data that are stored in multiple, autonomous and heterogeneous storage sources.

In data integration, systems are integrated if they act the same, look the same and they consume and or produce the same data. With the scale of data integration being done at present, very few are paying enough attention to one of the greatest threats in the IT dependent enterprise, namely security. Originally the database administrator had one database to protect. They had a handful of known users and defined access points. With data integration, database security has taken on a whole new dimension. There are now 'anonymous' users that need access at all times of the day and night. They need access from geographically dispersed locations. The database administrator's headache's twin brother, integrity, also takes on a scope to dwarf any diligent administrator.

Integration is the elusive higher state that organizations want to achieve, but end up paying inadequate attention to due to the difficulties in achieving true data integration.

Having one database to maintain and 'protect' was easy, relatively. Merging your data with an 'unknown' database's data creates integrity and security concerns that should never be taken lightly. In addition to these are the threats posed by transmitting data over public networks. Performing data integration opens a system up to several areas of threats (figure 1) that one must always be cognizant of. These threats may be identified by the following questions:

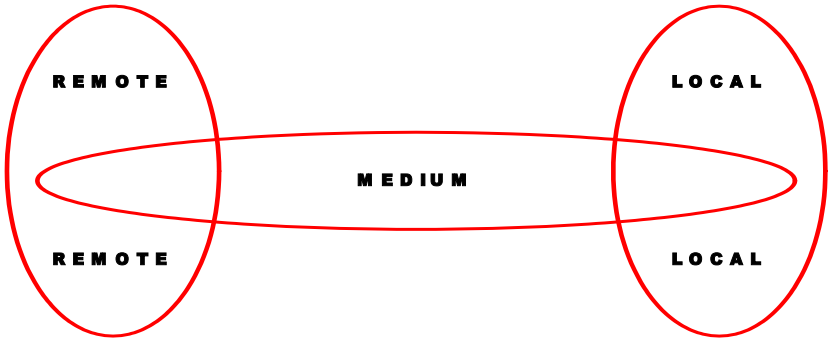

Figure 1: Areas of threat.

- Who controls the remote data?
- Is the source reliable?
- How do you enforce access?
- Is the data clean?
- How do we ensure the integrity of the data?
- Can we ensure the secure and reliable transmission of the data throughout?

The questions as raised above, such as source, medium and storage concerns that are raised in one form of integration, are also valid in every other form of integration. The risks are elevated in an integration process, as there are so many more factors to consider. These include but are not confined to the practice of having unencrypted data residing in vulnerable locations, inadequate user authentication and authorization and the ever prevalent human error due to mismanagement of guidelines and policies.

The data integration solutions provided by the numerous vendor options available, tackle security as an 'add on' when achieving the overall objective, integration. Yet the underlying problem persists. The integration solutions have become efficient at achieving data integration overall, but, if one cannot guarantee the validity and integrity of the data, the 'single version of the truth', the reliability and accuracy of the results are suspect. The old saying, 'garbage in, garbage out' takes on a whole new meaning!

3 DATA INTEGRATION SOLUTIONS

As mentioned, data integration attempts to provide a unified view of the information available. Data integration provides a high performance and cost-effective method for companies to integrate global data from remote locations as well as share data with partners across their extended enterprise.

The traditional data integration tools provide extract, transformation, and load (ETL) capabilities but fall short of creating an enterprise-wide awareness of the available knowledge resources. We therefore see specialized areas of integration such as Enterprise Application Integration (EAI), Distributed Data Integration (DDI), Point-to-Point Integration, Data Warehousing (DW), Data grids and Federated Databases to mention a few. One of the more prevalent types of integration today is EAI.

3.1 Enterprise Application Integration

Enterprise application integration is the process whereby data and business processes are shared by the data sources and applications in an organization. Due to its scope it is difficult to define the exact meaning of EAI. Buyens defines it as "...the ongoing *process* of putting an *infrastructure* in place, so that a *logical environment* is created that allows *business people* to easily deploy *new or changing business processes* that rely on information technology" (Buyens, 1999).

Due to the continuous changes in the information technology environment, it is necessary to adapt to or change the security configuration when using enterprise application integration. In certain instances EAI could expose the organizations to further risks as it creates a multitude of access points. This can be due to the frequently changing implementations of EAI, as well as the various components that need to be changed.

The vulnerabilities of EAI can be grouped into the following areas (Varlow, 2002):

- Data traversing the network between the applications can be disclosed.
- Information within the EAI environment can be improperly altered.
- Loss of system availability due to denial of service.
- The incorrect execution of business processes.

3.2 Point-to-Point Integration

Point-to-point integration systems are linked applications and databases that have been connected by means of hand-coded, proprietary connectivity systems. Point-to-point integration is an effort to address an isolated business need, and must therefore not be seen as a solution for cost-effective reuse in other business areas.

Having a patchwork of point-to-point links is a weak foundation for a strategic multi-source data integration system (Routly, Pottas, 2003). The reason for this is that it requires regular and

substantial investment in the technology while at the same time only delivering a limited value to the company. One of the issues that can come to the fore when using this approach, is the aspect of security. Authentication especially is often not adequately addressed. It can also be very expensive to implement, as a quick fix solution is not reusable. Every time that something changes, be it on the source or destination side, additional code often has to be added. Links in the framework have a tendency to increase exponentially over time. This is time-consuming, as there are continuous additions and deletions that need to be tested, maintained and secured.

One of the more common forms of data integration is data warehousing.

3.3 Data Warehousing

The data warehouse is a conglomeration of enterprises' transactional data. The goal of the data warehouse is to be able to allow users to perform queries and extract reports from the underlying data (Kimball, Ross, 2002). The data that resides in the warehouse is specifically optimized for performance and ease of use.

The movement and storage of large quantities of data as in this integration genre, pose significant risks. For example, it is not uncommon that end users have the ability to create their own queries. This results in applications that are created within a short time frame, creating security risks due to the ad hoc nature with which they are assembled. As with pivot table solutions, the end user has a great amount of flexibility when customizing their solution to meet their needs. Thus there is a distinct problem when enforcing security at the users' end. The proliferation of solutions has given rise to a variety of products, that do not have security in mind, but rather the ease and flexibility of extracting an answer.

The most critical area of the data warehouse is the ETL process. As in all integration projects the data source poses a significant risk. During the ETL process the integration system requires access to the source via some form of user ID and password mechanism. As in all development environments the ETL developers should not be exposed to more data than they require in order to achieve their purpose.

One area of data integration that is becoming more prevalent due to its scalability and adaptability is that of federated databases.

3.4 Federated Databases

A federated database system is a collection of cooperating but autonomous component database systems (Sheth, Larson, 1990). Therefore, the term federation refers to the cooperation of independent systems and in the field of databases; it is reflected by controlled and limited integration of autonomous databases. The information used is distributed across each of the databases in the union.

The primary concerns in federated databases are that of autonomy and heterogeneity. Autonomy refers to the amount of control the local database system has over the system. This includes the joining to or disjoining from the federated union. The heterogeneity refers to the multitude of schemas that need to be integrated with one another. This integration allows the user to access and manipulate data on multiple databases as though working on a single database transparently. Each database in the union may have its own security and access control methods. Once in a union there are the local security and access control policies as well as the policies as governed by the federated security policies (Tari, Fernandez, 2004). This requires that during an

access, the federated server must not violate the participating servers' security policies. In addition to the heterogeneity of the schemas, there is also the heterogeneity of the participating server's security policies. It is clear that the merging of policies from multiple servers is not a moot point. One must bear in mind that semantic heterogeneity also plays a role, as there could be a difference in how data is classified in different organizations (Tari, Fernandez, 2004).

The problem in federated databases is that one has a multitude of users that have varying levels of rights, accessing a collection of databases that have varying degrees of sensitive data. As mentioned earlier, the varying access rights of a user from one database to another raises complexities. It is therefore advisable that security must be enforced at the federated level. The problem that persists is that although there are efforts to provide for secure cooperation between databases (Sheth, Larson, 1990) (Jonscher, Dittrich, 1994), the solutions are geared towards a specific database model or group of models.

Data integration is a complex procedure for organizations where one specific approach does not apply to all problems they have. Which solution is best will depend on the specific requirements and the problems being solved.

4 INTEGRATIONS SECURITY PARADIGM

A security solution should not complicate the primary objective of integration by implementing complex security procedures and mechanisms, but should have a uniform method of safeguarding an integration transaction *at all times*. This includes being able to ensure the integrity of the cycle the data passes through, and the data itself before, during and after integration. Secure data integration with valid integrity is the objective.

Conventional integration projects merge millions of rows and terabytes of data. A security solution must be the hidden guard, one that only intervenes when needed and at no other time. Working with data in an organization, even without a data integration project, is time-consuming. The effort required to maintain, manage and protect the data is extensive – one may get the impression that the data itself is the enemy. When turning this data into extractable value, the way we secure and validate this data, requires a new way of thinking in its implementation.

With all this said, there is still a distinctive need for a security platform on which to base secure data integration solutions. The platform should encompass all security facets that could be required in the integration environment. A useful starting point from which to approach this would be the security services in ISO 7498 part two. ISO 7498 has been superseded by ISO 10745 and ITU-T X.803, ISO/IEC 13594, ITU-T X.802, ISO/IEC 10181-1 and ITU-T X.810. This does by no means imply that they are no longer valid, the document has been superseded, not the concepts.

The common thread that runs through all forms of data integration are entrenched in three areas of commonality. These areas are the data source, the network medium and the data store. In creating a solution that would ensure *end-to-end* security in a distributed data integration scenario, an analysis of the defined areas would be required. The categorization of security services as found in the ISO 7498-2 provides a useful framework for such an analysis.

5 AN ISO 7498-2 PERSPECTIVE

5.1 The ISO 7498-2 Security Elements

The ISO 7498-2 standard provides a reference model for communications between open systems. It also specifies a set of security services that can be used as a frame of reference for exploring the services required in distributed data integration. When viewed in the form of a control matrix (refer to Table 1), the services in the model can be shown as to their implementation in the main areas of integration.

Security Element	Data Store	Network Medium	Data Source
Confidentiality	√	√	√
Data Integrity	√	√	√
Authentication	√		√
Non-Repudiation	√		√
Access Control	√		√

Table 1: Security services per area.

When applied to data integration, the various services could be interpreted as (ISO, 1984):

Confidentiality: The property that provides protection to ensure that information is not made available or disclosed to unauthorized individuals. It should also allow for the prevention of exposure of the data due to the analysis of the traffic flows

Data Integrity: The property that provides for assurance that the data has not been altered or destroyed by an unauthorized individual during the integration process.

Authentication: The property that provides for the authentication of an individual for example, the individual is validated as the person who they claim to be. In addition it provides for origin authentication, which proves that data originated from a data source that is verifiable.

Non-repudiation: This protects against one of the individuals or systems involved in a communication denying that the data was sent or received by them.

Access Control: It protects against the unauthorised use of a resource by a user, which for example, could be the data store or the accessing of a data source.

The intersections of Table 1, which provide a mapping between each of the security elements and the three generalized integration areas, are now investigated according to threats and possible control measures.

5.2 Data Source Issues

In data integration, as in most forms of data exchange common elements come to the fore. The authentication, authorization and validation of users and systems are paramount. Thereby access to resources and data can be controlled.

5.2.1 Non-Repudiation

Any system used in an integration procedure should not be able to deny participation, thereby defining the universe of possible members of source data. The use of digital signatures can feature predominantly in this area. A data source that signs any data that has been requested cannot easily refute that the data originated from its system. The digital certificate of a system or user is proof of the identification of that said user.

5.2.2 Authentication & Access Control

Any system or user accessing the source systems will need to be authenticated against a group of users and privileges. The privileges assigned can be accomplished by assigning a specific user rights or the rights can be assigned to a role or a group. Authentication can be achieved by a challenge/response type of scenario with a username and password featuring predominantly. The goal of such mechanisms is to prevent unauthorized access to the resource, by validating the user and verifying that they are who they claim to be. Authentication and access control must be enforced and maintained throughout the data access period.

5.2.3 Authenticity

When using a digital signature we can be assured of the origin of the transmission and that the data remained unaltered throughout the transmission. The digital signature does not assist in or contribute to confidentiality, which is the role of encryption. The message that is sent has a message digest generated of it. The sender then signs the plaintext with their private key before sending it. At the receiving side the recipient creates a digest of the plaintext and compares it to the digest that the sender signed and if they compare, it signifies that the message has been unaltered. What this equates to is that the data sender is validated by the recipient by their public key and the message is proved to be unaltered.

5.2.4 Confidentiality & Integrity

When extracting data from an independent source the concepts of privacy, confidentiality and integrity must be kept in mind. The integrity of the sources data must be defined before transmission so that confirming its integrity post transmission can be easily ascertained. Ways of achieving this can be by providing message digests of the data before transmission, then affixing this to the respective segments of data. In order to ensure that the segments are unaltered, they can be digitally signed as an additional level of verification. It is also important that the privileges granted to accessing users are thoroughly controlled so that access is only granted to a specific file, database or segments thereof. This allows the confidentiality of data to be maintained as only valid users can then access it.

Stationary data poses a significantly greater risk to being compromised than flowing data. Should it be required that the data be stored, once the data has been extracted, verified and digitally signed and it is not immediately transferred to the central storage location, it should be encrypted whilst in stasis. Encrypting the data whilst in temporary storage assists in maintaining confidentiality of the data.

5.3 Integration Path Issues

Mergers are more prevalent in the current IT environment. It is therefore understandable that the various institutions would need to communicate and exchange data. As in distributed data integration, the various data sources would be dispersed among branches on different continents. Therefore there would be a need for these sources and stores to exchange information in a secure manner that would not impinge on the integrity of the data found in the systems. It cannot always be assumed that the data paths used will take place on leased lines or private networks. It is therefore becoming more prevalent that data transfers should be encrypted. If we were to imagine these transactions from the angle of the adage, “garbage in equal’s garbage out”, by enforcing data integrity in transactions, we are keeping the garbage out. By ensuring data integrity in the medium, we can ensure its quality, the very quality that an organization depends upon in order to make informed decisions.

The network is a crucial entity in organizations’ communication processors. One must always be cognizant of the fact that all transmissions can be intercepted. If one can ensure some form of transmission security, this will then translate into a secure network. Thus the use of encryption plays a role in ensuring the security of the data during transfer and more commonly while in stasis.

5.3.1 Confidentiality & Integrity

The objective when using encryption is that we must ensure some form of confidentiality, that is the message must remain a secret. Using encryption during integration protects the data during transmission when taking place over an unprotected medium such as the Internet. Encryption assists in providing the data with a form of integrity so as to ensure that the data remains unaltered during the transmission between source and storage systems.

Encryption is important for information security as it ensures the confidentiality of the message and that the message is valid from an integrity and authentication standpoint. It must be stressed that with the above mechanisms in place it can be stated that the message is "safe" whether the medium itself is secure or insecure.

Ideally one would like to ensure the safety and security of the medium between source and storage systems. But, when performing data integration between geographically dispersed systems it is not always possible to guarantee each segment in the data path as segments and internetworking devices fall outside of the organization's control. In addition traffic on public networks is susceptible to interception and alteration thus negating the very integrity of the data. By ensuring a level of confidentiality and integrity in the medium, the security of the medium is enhanced.

5.4 Data Store Issues

In distributed data integration implementations the data store is separate from the data source often by a substantial geographical distance. In other types of data integration the source and store could be mere meters apart. Therefore the area where the integrated data will reside is viewed as a separate entity. Several processes could occur the moment the data enters the data store network.

5.4.1 Auditability

Each process throughout the data transport lifespan should be logged; thereby providing for auditability of all data transactions and movements, occurring in the system. Karger emphasizes the greater use of logs in security enforcement (Karger, 1988). He also goes on to state that such logs should be available to access control software. This will enable them to be used in the enforcement

of access control decisions as well as a historical means of determining of who has access to what in the system. The logs provide an additional layer of protection by listing successful and unsuccessful access to data.

User authentication in conjunction with logs can ensure that data privacy is maintained, by ensuring not only that valid users only access the defined data to which they have rights to, but in addition listing those users who have valid access but are in effect viewing confidential information.

A central, secure logging server allows for a greater amount of historical data to be retained. Intrusion detection systems play a significant role in that unauthorized traffic on, and access to the network can be timelessly detected thus minimizing risk to an exposure of the data. Using IDS in conjunction with a central logging system allows for trend analysis of the valid transactions occurring within the network, thereby allowing for the isolation and identification of the suspicious activities that are occurring on the network.

5.4.2 Confidentiality

In most integration implementations during the traversal of a WAN the data is encrypted, but as soon as it enters the data store network the encryption is removed. This leaves the data in a plain text format during its time spent from the data store firewall to the store database. In itself this is a threat to the security and integrity of the data as it exposes the data to untold threats. It goes without question then that the data must be protected throughout the integration process.

The key concept is to ensure adequate workarounds should the data need to traverse devises that use caching and switching functions that would not be able to function as encryption would make the operation of these functions improbable. The use of symmetric encryption is an option as it is faster to encrypt and or decrypt than asymmetric encryption. If the data is in an XML format, XML encryption can be used.

5.4.3 Access control & Authentication

The ability to make use of the data store should be based on access lists maintained by the system or some trusted third party. Access to resources on the system must be controlled by means of a digital certificate and an additional form of authentication. The importance then lies in the means to associate an individual with a digital certificate so that they cannot be denied and the ownership cannot be refuted. The identification of the individual accessing the resource must be achieved by the individual being able to produce some information certifying their identity, thus the use of digital certificates. The systems ability to identify an individual is based upon the assumption that the system has the user records and rights or a trusted third party stores them for the system. This gives rise to the possibility that the user listing and corresponding rights can be stored locally or remote to the system but are accessible by the trusted third party.

The authentication of individuals should be the responsibility of a trusted central facility. This facility is entrusted with the verification of user's identities and the rights associated with them. It is also tasked with the enforcement of such rights with respect to the data being accessed. The importance of access control is increased as with the integrated data, there could be new users that require access to the combined data.

5.4.4 Data integrity

When a data is received a message digest of the data is created to ensure that it matches the digest of the data that is included in the file, thus ensuring the integrity of the data.

When inserting data into the source system; data can be flagged with markers that will assist in the decision making process made on the said data. This will allow for the data to be used, but will allow for the reliability to be graded. Ammann and Jajodia mention the use of *integrity markers* that give values such as correct, acceptable, and wrong but usable and so forth to mark data (Ammann, Jajodia, 1997). This will allow the results derived from the data to be graded as to their reliability and the level of dependence that can be placed on them when making decisions.

It is possible that there could be an integrity issue with data from an independent store; therefore the above mechanisms assist in the overall integrity of the data to be ascertained. If the data is deemed trustworthy after a period of time, its "grading" can be altered to reflect its new status. One must not overemphasize the integrity aspects, but integrity checking is not only about the checking of the database itself but in the ensuring that any changes to the data do not affect the consistent state in which the database resides (Eaglestone, Ridley, 2002).

The security of the data store should be the most important area when implementing security in data integration. It is the origin of the integration process; it is the place to which any data that results from the integration process will return to. It is the store of the data and it is the area user's access in order to retrieve the information they need.

6 CONCLUSION

As discussed in the paper, the process of data integration requires a renewed focus with regards to ensuring security. Implementing security in a single environment requires clear definitions and solutions. In data integration we could possibly be presented with several remote environments in which we would have to ensure security due to the heterogeneity of the types of implementations possible.

Existing technologies can be used towards achieving the paradigm shift that is required. If the importance of data and its quality are becoming such well-known concepts, surely an approach to safeguarding this asset must be implemented in order to gain the level of security and integrity that the modern day organization demands from their integration solutions.

7 REFERENCES

Ammann, P., Jajodia, S. (1997). Rethinking Integrity. *IEE Concurrency*, 5(4), 5-6

Buyens, M. (1999, November). Enterprise Application Integration (EAI). *ID-SIDE*, 11

Eaglestone, B., Ridley, M. (2002). Verification, Validation and Integrity Issues in Expert and Database Systems: The Database Perspective. In the 9th International Workshop on Database and Expert Systems Applications (DEXA'98), 22

Eastwood, G. (2001). Data is the key. Retrieved 19th December 2003 from Managing Information Strategies Web site http://www.misweb.com/printmagazine.asp? doc_id=21886

Halevy, A. Y. (2001). Answering queries using views: A survey. *Very Large Database Journal*, 10 (4), 270-294.

Ibrahim, H. (2002). A Strategy for Semantic Integrity Checking in Distributed Databases. *In Proceedings of the Ninth International Conference on Parallel and Distributed Systems*.2002, 139

ISO (1984). ISO 7498, *Information Processing Systems - Open System Interconnection- Basic Reference Model*, International Standards Organization,Geneva, 1984.

Jonscher, D., Dittrich, K.R. (1994). An Approach For Building Secure Database Federations. *In proceedings of the International Conference on Very Large Databases, Santiago*, 24-35.

Karger, P. A. (1988). Implementing Commercial Data Integrity with Secure Capabilities. *In Proceedings of the 1988 IEEE Symposium on Security and Privacy*, 1988, 130-139.

Kimball, R., Ross, M. (2002). *The Data Warehouse Toolkit*. John Wiley & Sons, Chichester, 2002.

Routly, W.A., Pottas, D. (2003). HEDIN: A Model to Facilitate Heterogeneous Data Integration. *Mini Thesis in fulfillment of Bachelors Degree in Information Technology, Port Elizabeth Technikon*. Port Elizabeth

Sheth, A.P, Larson, J.A. (1990). Federated database systems for managing distributed heterogeneous and autonomous databases. *ACM Computing Surveys*, 22(4), 183-236.

Tari, Z., Fernandez, G. (2004). Security Enforcement in the DOK Federated Database System. *In Proceedings of the tenth annual IFIP TC11/WG11.3 International conference on Database Security,* 1997.

Varlow, S. (2002). Security Strategies for EAI. *EAI Journal*, September 2002, 42-44.