Secure Data Integration Systems

Fatimah Y. Akeel[1,2], Gary B. Wills[1], Andrew M. Gravell[1]
[1]*Electronics and Computer Science, University of Southampton, Southampton, U.K.*
[2]*King Saud University, Riyadh, Saudi Arabia*

Abstract: With the web witnessing an immense shift towards publishing data, integrating data from diverse sources that have heterogeneous security and privacy levels and varying in trust becomes even more challenging. In a Data Integration System (DIS) that integrates confidential data in critical domains to contain a problem and make faster and reliable decisions, there is a need to integrate multiple data sources while maintaining the security levels and privacy requirements of each data source before and during the integration. This situation becomes even more challenging when using cloud services and third parties in achieving any part of the integration. Therefore, such systems face a threat of data leakage that compromises data confidentiality and privacy. The lack of literature addressing security in DIS encourages this research to provide a data leakage prevention framework that focuses on the level prior to the actual data integration, which is the analysis and early design of the system. As a result, we constructed SecureDIS, an architectural framework that consists of several components containing guidelines to build secure DIS. The framework was confirmed by 16 experts in the field and it is currently being prepared to be applied on a real-life data integration context such as the cloud context.

1 RESEARCH PROBLEM

Integrating personal or sensitive data sources originating from different organisations that vary in security and privacy requirements is a challenging task. The main reason for this is that the integration occurs at two different levels, one is the data level and the other is the level of the security and privacy requirements that belong to each data source. The latter raises concerns of conflict between security and privacy requirements (Yau and Chen 2008). In addition, there is an issue of difficulty in maintaining those requirements throughout the complete integration process. To further aggravate the situation, the entities providing the information, or participating in the integration, can vary in their levels of trustworthiness. Hence, the integration process should not be focused on the data level only. It should address the level of combining security and privacy requirements and consider trustworthiness.

Achieving data integration without considering maintaining the combination of the Security, Privacy, and Trust (SPT) aspects of the entities participating in the integration process leads to different threats. One important threat is *unauthorised access* to data, caused by weakness, mis-configuration, or inappropriate access controls models (Braghin et al. 2003; Watson 2007; Pistoia et al. 2007). Another example is the wide spectrum of inference attacks occurring from failure to address privacy (Fung et al. 2012; Boyens et al. 2004; Whang and Garcia-Molina 2012; Clifton et al. 2004). Yet another example is the untrustworthy behaviour caused by entities involved in the integration process, such as initiating transitive trust with other entities without the consent of the DIS (Fung et al. 2012). These threats are combined under a generic threat called *Data Leakage,* which is defined as the disclosure of confidential information to unauthorised entities intentionally or unintentionally (CWE 2013).

As mentioned, the failure to combine the SPT together in a system may allow data leakage to occur. Additionally, the mis-design of the SPT features of the system can lead, eventually, to data leakage. Therefore, data leakage threats can be controlled if the systems are designed to address SPT from the start and consider the combination of the SPT of each source and entity.

Current approaches found in the literature to secure a DIS in general, are either focused on a specific component of the system, such as the integration approach, or focused on a specific attribute over the whole system, such as privacy. However, there is a need for an approach that considers all the components of the system at the same time as to considering several attributes that contribute to the overall security of the system.

Akeel F., B. Wills G. and Gravell A..
Secure Data Integration Systems.

The data integration literature found covers the aspect of privacy in a specific component of the DIS, addressing privacy-preserving data integration and data mining techniques extensively. However, in terms of security requirements, there is a separate body of literature that is concerned with combining the security polices of entities collaborating together, and not particularly in data integration context(Cruz et al. 2008). Very limited literature has been found that discusses these two levels together, such as the work by Haddad, Hacid & Laurini (2012) and to the best of my knowledge, no literature has addressed the combination of SPT in a DIS context. In any case, many approaches simply assume that the entities collaborating or integrating are trustworthy.

This lack of literature encourages this research to investigate the security in data integration contexts, and to focus on the level prior to the actual data integration, which is the analysis and early design of the system.

2 OUTLINE OF OBJECTIVES

The main objective of this study is to find a solution that allows integration, collaboration, and data sharing between different organisations while maintaining individual security policies of the participating entities. We argue that maintaining the confidentiality and protecting privacy while considering trust in each entity in the DIS with a middle layer (e.g. cloud) will reduce the threat of data leakage.

Our approach focuses on guiding the design of DIS to include confidentiality, privacy, and trust aspects. This can be achieved by the following stages:

- Identifying the architectural components of DIS with a middle layer.
- Identifying the possible data leakage locations within the DIS architecture.
- Identifying the confidentiality, privacy and trust weaknesses that cause data leakage threats in DIS components.
- Creating a framework that contains the DIS architectural components.
- Creating an initial set of guidelines that aim to reduce possible data leakage threats.
- Linking the guidelines with the appropriate framework components.
- Confirming the framework and its guidelines with experts in the field.
- Using the framework on a cloud-computing context to assess its usefulness in reducing threats of data leakage.

3 STATE OF THE ART

This section provides the scope of the topics covered by this study and defines the key concepts. In addition, it discusses the themes of the reviewed literature and provides a critique relevant to the scope of this study.

3.1 Scope and Definitions

Data integration systems are usually complex (Russom 2008) and have different variations and forms. Therefore, to manage this study, the scope is focused on DIS that use a middle layer to manage the interaction between data sources and data consumers and achieve integration. The data sources used in such systems originate from different organisations and hence they vary in security and privacy requirements and trust levels.

The important aspects of the scope are defined as follows:

Security: is usually defined as the combination of confidentiality, integrity and availability (ISO 2014). The attribute of concern in this study is confidentiality; other attributes are assumed to be implemented. Confidentiality is achieved by limiting access to data to aauthorised individuals, entities and processes.

Privacy: is concerned with protecting personal information (Gollmann 2006) and determining when, how and what type of information can be exposed to others (Jawad et al. 2013). The attribute of concern is anonymity, to ensure that personal information is not exposed.

Trust: is the belief that an entity will behave in a predictable manner by following a security policy (Ross et al. 2014).

Data leakage: is disclosing private information intentionally or unintentionally to unauthorised parties (CWE 2013).

3.2 Securing DIS Components

Few works in the published literature have suggested securing DIS as a whole by considering privacy and trust. However, trust is still an issue in distributed systems. Prakash & Darbari (2012) discuss several security approaches that aim to enforce trust, such as the use of trust models. They also discussed risk management as a method to evaluate trust. Van Den

Braak et al. (2012) propose a framework that aims to support data sharing among different public organisations. It preserves privacy through sharing data based on a need-to-know principle, where data is provided only when required for a specific process. The authors propose the notion of a Trusted Third Party (TTP). The TTP is responsible for integrating and sharing data between different government organisations. The proposed framework contains two parts: the first part is data integration techniques to achieve privacy, while the second part provides guidelines on data sharing that ensure security and trust. Nevertheless, the guidelines provided mainly focus on the *Integration Location* and *Data and Data sources* components of the DIS rather than the system as a whole. However, the integration covered was across government organisations, and thus, still under the same security policies; therefore, the risks of violating the integrated security policy were not present as one of the security and privacy challenges that the approach overcomes. In addition, in this approach, trust is assumed, and it lacks guidance on the need to establish trust or evaluate it in relation to other entities. Therefore, when security and privacy challenges are discussed in this work, they do not include trust threats because it is assumed in the first place, and those challenges are not particularly addressed as data leakage threats. Finally, although the proposed guidelines are reasonable to prevent data leakage threats, they are not linked to any specific phase of the software development and therefore it is not clear when to apply the guidelines to the SDLC.

The approach proposed by Clifton et al. (2004) is a privacy framework for data integration. The purpose of the framework is to provide an insight into the privacy challenges in data integration. It provides several research directions, one of them emphasizes the need for a privacy framework that considers users privacy views to expose and hide sensitive attributes, privacy policies implementing these views and a purpose statement specifying which data is allowed to be accessed and integrated. The solutions discussed to preserve privacy in data integration consider the following components: *data and data sources*, *integration approach*, *data consumers* and *security policy* only. The integration location and the management of the process of the integration are not addressed within the framework. In addition, it is not presented with any link to software development. It should be mentioned that the authors have addressed data leakage mainly through discussing the difficulty of preventing multiple query attacks. The heterogeneity in the security and privacy policies are only briefly addressed and there is no specific focus on them in terms of their relation to the *Integration Approach*. Trust, on the other hand, is not addressed at all as one of the challenges within the framework.

The work of Bhowmick et al. (2006) is very similar to that of Clifton et al. (2004); however, it proposes a more detailed architecture and a framework for privacy-preserving data integration and sharing deployable DIS. It includes security and policy considerations by suggesting adding a security policy component to the system. It also provides several suggestions on preserving privacy in different DIS components. The architecture covers most of the DIS architectural components except the management. However, the level of detail provided in terms of integrating the various security policies of the data sources and the integration location is not sufficient. In addition, the suggestions provided are not listed in the form of technical or practical security guidelines. It also does not present the framework from a clear specific development phase.

A policy integration method that combines the authorization policies of data sources integrating and sharing data is proposed by Haddad, Hacid, & Laurini (2012).The method focuses on creating a global security policy that consists of local security policies of the participating data sources in the integration process. This global policy is enforced within the mediator level i.e. *Integration Location* component. According to this approach, queries will not be processed unless they are authorised by a source. Therefore, the access to sources will be preserved. One of the limitations of this approach is that it covers the security policies generated by *Data Sources* and the *Integration Location*; but it does not consider the actual data *Integration Approach* and the *Data Consumers,* nor *Management*. In addition, it does not consider the trustworthiness of the entities participating in the integration process. Furthermore, it does not explicitly specify the software development phase in which the approach can be used.

Another work by Jurczyk and Xiong (2008) focuses on privacy preserving data integration. It proposes several protocols for data anonymization in addition, to a general architecture for data integration. Hu and Yang (2011) propose a privacy protection model for DIS by using semantic approaches. The works reviewed in this section are primarily focused on privacy, which can be related to security, and little or no attention to trust is given

in these works. In addition, the approaches provided are applicable to relational databases and do not consider other data formats.

Generally, the literature provides practical and applicable solutions expected to solve problems in a specific DIS component. However, the security of the whole system is discussed only in a limited way in the literature. Studies usually assume that the provided data is secure and comes from trustworthy entities. However, from a security perspective, data sources are considered an important and effective element to guarantee the security of a DIS. Data sources can therefore fall under the data-centric security category within the information security field, and having security-meta data would help in distinguishing secure sources from unsecure ones.

In organisations that employ data integration to integrate private data, there is a need to manage data access and authorization. A carefully selected access control model that enforces security policies is essential. Therefore, organisations need to create well-defined security policy that enforces data security, privacy and protection. To ensure the security of a DIS, the combination of the individual security policies of each data source needs to be carefully considered. There are many studies of security policies and access controls that cover policy integration in different contexts; however, there is an evident lack in considering trust. One possible reason is that organisations usually integrate data coming from their own data sources, which are assumed to be trustworthy.

In DIS, the resulted integrated data are normally requested using queries by data consumers, which can be humans or services. Data consumers are an important component of any DIS, as they can be a source of security and privacy violation. Many attacks on information systems, including DIS, are caused by data consumers, such as SQL-Injections to gain access to data that they are not authorised to. In the DIS context specifically, inference attacks and attribute correlations that lead to data leakage threats are usually carried out by data consumers. In addition, the consumer use of access control models that are not well evaluated leads to unauthorised data access. Therefore, it is important to consider data consumers from a threat point of view and plan to build the system in a way that prevents such attacks. The literature on DIS threats and attacks focuses on privacy attacks conducted by data consumers; other attacks are not discussed as they do not differ fundamentally from any other web-based applications attacks.

3.3 Integration Borders

This theme relates to the differences between integrating data within or outside an organisation and the effect this has on the security, privacy and trust of the data integration process.

3.3.1 Integrating Data within the Organisation

There are several domains where data is requested and integrated within the border of a single organisation or within a range of similar organisations belonging to the same sector. Enterprise Information Integration (EII) (Halevy et al. 2006), healthcare (Bhowmick et al. 2006) and, scientific research (Ray et al. 2009) are all examples of where this type of integration can occur.

There is a large volume of published studies on data integration that takes healthcare domains as a context, such as the works by Boyens, Krishnan & Padman(2004),Tian, Song & Huh (2011), and Eze, Kuziemsky, Peyton, *et al.*(2010). There are also several studies concerned with security and privacy issues in healthcare in general, for example the one by Meingast, Roosta & Sastry(2006). This large body of existing work makes healthcare approaches on security and privacy useful to survey. Healthcare is a unique sector, with characteristics that are not found in other sectors (Avison and Young 2007).

This means legislation and policies exist that strive to protect this kind of domain to maintain data integrity and confidentiality. In the UK, healthcare organisations have to comply with the Data Protection Act (Philip Coppel Qc 2012),whereas, in the United States healthcare organisations follow the HIPPA act (U.S. HHS 1996).

In many healthcare organisations, a DIS is systematically monitored for compliance with legislation and policy as well as other criteria (Eze et al. 2010). Reviewing 20 of 30 Health On the Net Foundation Code of Conduct (HONcode) accredited American online healthcare appointment websites for compliance with basic principles of security and privacy, Hong, Patrick, & Gillis (2008) found that only 8 of the 20 websites are secure and 12 of the 30 were not showing privacy notices to patients on their websites. They found that there is a gap between the ideal security and privacy requirements and the reality in applying them. They therefore, propose several steps to overcome this gap and make it possible for healthcare organisations to be compliant with legislation and security guidelines.

There are several requirements needed in the

healthcare domain. One is to balance security and interoperability between healthcare actors and organisations. Dawson, Qian, & Samarati (2000) suggest an approach that allows multilevel secure data sources to integrate and provide the results to external applications while maintaining their security levels. In addition to interoperability, Armellin et al. (2010) propose a system that publishes healthcare documents that provide interoperability and complies with privacy laws.

Another requirement is aiming to preserve privacy while integrating healthcare data. For example, building automatic data mashups that are aware of privacy concerns (Barhamgi, Benslimane, Ghedira, Tbahriti, et al. 2011; Barhamgi, Benslimane, Ghedira and Gancarski 2011). In addition, access controls used in healthcare systems can be extended to adapt to privacy requirements (Hung 2005).

3.3.2 Integrating Data outside the Organisation

Integrating data outside of the organisation means that the integration location or part of it is outside the organisation boundaries, for example, when data sources are handled by cloud services and/or third parties. In addition, the data sources may reside outside the organisations boundaries. The following subsections discuss each of these cases.

3.3.2.1 Using the Cloud as an Integration Location

Clouds suffer from many security, privacy and trust issues and therefore they need to comply with regulatory laws for data protection (Takabi et al. 2010; Youssef and Alageel 2012). In addition, to prevent unauthorised access, they need to deal with the heterogeneity of security components and multi-tenancy (Takabi et al. 2010). The fact that the clouds are not under an organisation's physical control elevates the problem of managing data security (Reeve 2013), especially when there are no standard rules and regulations to deploying the cloud (Saeed et al. 2014). These aspects should be considered in any data integration security model, to emphasize the importance of investigating the location of integration.

3.3.2.2 Using Third Parties

Third parties are used in data integration applications for different purposes. On one hand, organisations may want to outsource the data to a third party to analyse it and find out aggregation statistics (Xiong et al. 2007). Alternatively, an organisation may require an entity to handle access control to personal integrated data, such as the approach described by Van Den Braak et al. (2012) which uses a trusted third party to handle access controls to government data in the public sector. The proposed approach uses privacy preserving data integration and collaboration.

Harris, Khan, Paul, & Thuraisingham (2007) argue that, in general, data integration applications that handle critical data, such as emergency response and healthcare awareness, need to share their data with different organisation s to make effective decisions. The authors discuss standard-based approaches to secure data across organisations covering different types of data and different types of domains. Their work emphasizes the need to enforce security policies and create standards or guidelines to govern application in critical domains.

3.3.3 Accessing Data outside the Organisation's Boundaries

There are cases when there is a need to integrate data from public data sources with an organisation's private data sources. This integration leads to different challenges, such as the lack of a form of privacy measurement, i.e. measuring the amount of privacy lost when data is exposed (Pon and Critchlow 2005). Another work, by Yau & Yin(2008) proposes a repository for data integration across data sharing services by collecting data based on user's requirements. If the repository is compromised, only the result of the integration is revealed and the original data will remain intact. The work by Mohammed, Fung, & Debbabi (2011) proposes two algorithms to overcome the challenges of revealing data coming from different data providers, using game theory.

3.3.4 Discussion

Integrating data within the organisation could be considered safe to an extent. The reason is that many of the entities involved in the integration process are within the organisation and are under the same security measures. The data sources are well known and the integration location is within the organisation's boundary. Therefore, the concerns about security are controllable, to some extent.

However, integrating data outside the organisation can be very critical. The reason is that many of the entities involved in the data integration process use different security models and have different privacy requirements. These entities can be

data providers or integration locations, which may not be secure enough.

Integrating data coming from outside the organisation raises issues on security policies. One aspect to consider in integrating security policies, especially when integrating data outside the organisation boundary, is to include the organisation's own policies with the external policies and the government legislation related to data protection, to ensure the security of the overall system.

Some organisations need to use trusted third parties to handle their data. One possible reason is that an organisation may have to respond to a significant number of requests and cannot respond in a timely manner. Another reason would be that an organisation may have lack of technical skills or infrastructure to handle the data. It therefore, transfers this responsibility to a third party to take over consumers' requests. As a result, releasing data to trusted third parties is critical, as the organisation may not monitor or track private data processing and movement exacerbating security concerns. In the real world, companies rely on legal agreements of data disclosure. Few use technical enforcement of data movement. However, the literature lacks coverage of this specific aspect in the data integration context. One study, by Van Den Braak, Choenni, Meijer, *et al.* (2012), proposes the use of trusted third parties; however, security concerns still arise.

3.4 Covering Security, Privacy and Trust

In studies that aim to secure DIS, the focus on the Security, Privacy and Trust (SPT) aspects varies. Some studies focus on SPT as separate aspects; other studies combine two of the SPT aspects. However, only a very limited number of studies have focused on SPT combination in a DIS context. The following section investigates these studies and how they focus on the aspects of SPT.

Secure data integration, mining, and sharing are addressed in the literature as approaches to SPT *separately*. In terms of ***security in DIS***, several recent studies, including the one by Haddad, Hacid, & Laurini (2012) and Begum, Thakur, & Patra (2010), have focused on security policy integration and conflict reconciliation and their uses to answer users' queries. Other studies have proposed extensions and improvement to RBAC to adapt to the integration context, such as the work by Lamb, Power, Walker, *et al.*(2006).

However, ***privacy in DIS*** has the lion's share of research. Privacy-preserving techniques are well established in the literature, spanning a range of different topics from privacy in peer-to-peer DIS to anonymization techniques. Bhowmick et al.(2006) propose a privacy preserving DIS framework that emphasizes the need to consider the balance between privacy and data sharing. This perspective has been later addressed by many studies: the work by Pasierb et al. (2011) presents different approaches to privacy-preserving data integration in e-healthcare systems, while the same concept applies to web services and data mashups by Barhamgi, Benslimane, Ghedira, *et al.* (2011b, 2011a).

Finally, in terms of ***trust in DIS***, there are many distributed trust models that can be adopted in DIS and can be used to determine the level of trustworthiness of either data providers or third parties. The work by Treglia & Park(2009) suggests a trust framework for intelligence information sharing between agencies. Other approaches focus on computational trust using either policy-based trust or reputation-based trust (Artz and Gil 2007)which can also be applied to DIS. Other studies acknowledge the need for a *combination between areas*, for example, the work by (Hung 2005) combines security and privacy by extending the RBAC model with privacy-based extensions.

Security combined with Privacy and Trust (SPT) has recently gained attention from the research community. Systems security is complicated and influenced by many other areas and therefore it cannot be addressed solely. Morton & Sasse (2012) supports this argument by proposing an integrated SPT framework to create an effective privacy practice in any information system. Considering human factors, another work in progress (Flechais et al. 2013) also takes this holistic perspective. It discusses authentication taking into account SPT in banking context in Saudi Arabia. Other studies such as that of Manan, Mubarak & Isa (2011) emphasize the need for such a research direction. In addition, a recent work on federated identity and access management created an access control solution that encompasses SPT, considered together (Khattak et al. 2012).

This concept is applied in a limited data integration context where Van Den Braak, Choenni, Meijer, *et al.* (2012) provide a framework for sharing and integrating data among public organisations in which ensuring security and privacy is achieved by using a trusted third party to manage access controls to private data.

However, to the best of my knowledge there is no single approach to preventing data leakage in DIS

that considers SPT aspects together, although they are very important to protect confidentiality and allow sharing data in a secure fashion.

3.5 Discussion of Data Leakage in DIS

Many vulnerabilities in software are caused by flawed design (McGraw 2004). Therefore, data leakage as a generic security threat can be caused by weaknesses in DIS design. The following subsections discuss these issues from security, privacy, and trust perspectives.

3.5.1 Design Issues Related to Security

Due to the heterogeneity and distribution of DIS components, security threats to DIS can arise from any component of the system. Threats can start from data sources that may not have adequate security and privacy meta-data, which define their level of sensitivity, and therefore, the DIS may face difficulties in maintaining the data sources' policies throughout the whole system. Moreover, different trust and security concerns may also be faced in the middle layers such as the cloud, where the data is mined, pre-processed, integrated, and prepared for presentation, in addition to many different tasks. Finally, data consumers where the data is accessed and queries are answered may pose threats as well.

Security in DIS is important, as it is to any information system. Having appropriate organisational security objectives and conducting early security analysis according to well-defined security requirements help in shaping any DIS towards providing a better security. Security is a comprehensive feature that includes many attributes; however, since many approaches proposed in a DIS context mainly focus on creating global security policies, enforcing security policies using access controls, and managing access to data this study will focus on the confidentiality as an attribute security. Some examples of how confidentiality in a DIS is applied: 1) Authorization to access data sources and the results of the integration is utilized by access control; 2) Data disclosure should not be disclosed equally to all user roles but should be limited to users with appropriate roles; 3) Considering security and privacy policies associated with data sources.

However, mis-configured access control models or selecting inappropriate ones can be a major design flaw that causes systems to be unsecure. It is important to adopt a suitable access control model to guarantee authorization to access data and guarantee its confidentiality.

Another issue that increases the possibility of data leakage is lack of knowledge of DIS users. Users' ignorance about legal issues in data management allows unauthorised access to occur(Batty et al. 2010). This can include developers and data managers as well. Therefore, a proper training is required to base the DIS design on updated knowledge of data management legal issues.

3.5.2 Design Issues Related to Privacy

Several design flaws that violate privacy cause data leakage. Firstly, the data sources used in the DIS may be originally predictable and very easy to link, or it may miss very important security meta-data that defines its level of sensitivity. Therefore, the DIS becomes vulnerable to different privacy attacks. A good design needs to create data sources that are very hard to link, or, in case of external data sources, it uses techniques that refine the data to minimize the ability to link information.

Secondly, although anonymization techniques are a popular solution to privacy, they are not always sufficient, for two reasons. On one hand, it is not the Personal Identifiable Information (PII) only that a DIS needs to worry about; it should worry about all the other attributes that cause inference attacks, such as Quasi Identifier (QID). On the other hand, a DIS needs to have customised anonymity levels that are suitable for the users accessing the data, as failing to manage anonymity discloses private data (Meingast et al. 2006).

Thirdly, systems that allow multiple consecutive queries to data sources (Clifton et al. 2004), especially sensitive data sources, are prone to privacy violation, as users can use the results for inference attacks, such as inferring conditional information from non-confidential data or from statistical aggregates, as mentioned by Zhang, Zeng, Wang, *et al.*(2011). Therefore, there is a need to manage the number of queries to private data sources, along with the users' roles and authorization level.

Considering these issues in DIS design will result in minimizing threats such as inference attacks, attribute correlations, and insiders attacks.

3.5.3 Design Issues Related to Trust

Designing systems that use external data sources coming from different organisations entails trusting the entities to provide accurate and reliable information. Using cloud services that process and integrate data, in addition to providing services to query and analyse the data (Carey et al. 2012) is

very risky. Clouds are security and privacy critical and they still require more effort to increase their reliability (Saeed et al. 2014). The risks arise from using multi-tenancy public clouds that share physical infrastructure with untrustworthy users can lead to different attacks, such as cross-virtual machine attacks (Ristenpart et al. 2009).

Trusting third parties to handle data in any form needs to be carefully considered. Computational trust does not differ conceptually from human trust. A trust in entities may be gained by their reputation or by certain actions they perform to obtain trust. Data leakage can occur from transitive trust where a trusted entity reveals sensitive data to other entities (Fung et al. 2012). Third party rights on the data need to be determined (Meingast et al. 2006), and many critical questions need to be answered during the design of DIS systems, such as: Do third parties have authority over the data similar to that of the data owner?

Data transfer to clouds and third parties needs to be based on trust (Saeed et al. 2014); hence trust should be added to the design process, through considering the entities that the system deals with and conducting risk assessment and risk mitigation. A DIS should be designed considering the collaboration among other entities and enforcing trust models between these entities to guarantee security.

In a DIS context, trust is an issue to consider for data sources and integration locations as well as third parties involved in the DIS. It also extends to data consumers, where granting roles to users depends on their level of trustworthiness. Therefore, trust is an important aspect in a DIS that cannot be neglected; however, it needs to be balanced with other properties to achieve secure and reliable systems.

SPT issues are exacerbated in a DIS context due to the complex and distributed nature of a DIS, especially across organisations. These issues, which are summarized and categorized in Table 1**,** contribute to the threat of data leakage in DIS. The full threat analysis is discussed in our previous work (Akeel et al. 2014).

Table 1: Summary of DIS Threats Causing Data Leakage.

Category	Threats and Concerns
Security	Unauthorised access caused by:
	1) Inappropriate access control model
	2) Access control weakness
	3) Mi-configured access control model
	Ignorance of legal issues on data management
	Inapplicable confidentiality on merged data
	Confidentiality violations due to inconsistent regulatory laws
	Leaking data to the platform
Privacy	Inference attack: attribute linkage
	Inference attack: linking data with QID
	Inference attack: interval disclosure
	Inference attack: gathering information about queries
	Inference attack: use of non-confidential information and statistical aggregates, namely record linkage
	Inference attack: consecutive queries attack
	Insiders attack: data providers inferring data
Trust	Using clouds to process data
	Third parties rights on data
	Third parties and transitive trust

4 METHODOLOGY

The results of reviewing the literature and realising the research gap and linking that to the study main objective, the following research questions are proposed:

RQ1: What is an Appropriate Framework to Reduce Data Leakage Threats in a DIS with a Middle Layer?

Focusing on the DIS architecture with a middle layer, this research question is mainly concerned with finding an appropriate framework that helps in reducing data leakage threats. The proposed framework aims to consist of the basic architectural component of a DIS with a middle layer.

RQ2: In the Proposed Framework, What Are the Confidentiality, Privacy and Trust Guidelines Used to Reduce Data Leakage Threats?

Understanding data leakage threats and locations in the context of a DIS helps in suggesting an appropriate set of confidentiality, privacy and trust guidelines that aim to reduce those threats. These guidelines can be included in the framework under each architectural component.

RQ3: How Can the Proposed Framework and Guidelines Be Used to Reduce the Threats of Data Leakage in a Real-life Scenario?

Using the framework and its guidelines in a real scenario will help in evaluating the framework from the practicality, applicability to context and usefulness to reduce data leakage threats.

Based on the previously mentioned research questions, Table 2 summarises the research activities relevant to answer each research question.

Table 2: Research Activities.

Research Activities	Purpose	Research Question
Literature review about DIS + Expert reviews 1	Confirming the components of SecureDIS	RQ1
Literature review about data leakage threats + Expert reviews 1 and 2	Identifying data leakage threats and their locations within the DIS architecture	RQ1
Literature review about reducing data leakage	Create the guidelines that aim to reduce data leakage threats	RQ1
The synthesis and analysis of the results of each step of the previous research activities	A proposed framework with architectural components containing a set of guidelines (SecureDIS)	RQ1
Experts reviews 2	Confirming the guidelines proposed by SecureDIS	RQ2
Expert reviews 2	To find out whether the proposed guidelines are comprehensive	RQ2
Expert reviews 2	To know whether the proposed guidelines are practical	RQ2
The synthesis and analysis of the results of each step of the previous research activities	The confirmed framework and guidelines (SecureDIS)	RQ2
A case study analysis	To find out the appropriate context to use the guidelines in	RQ3
A case study analysis + study of data leakage threats	To customise the guidelines to the appropriate context and apply them	RQ3
A case study + metrics	To measure reduction in data leakage	RQ3
The synthesis and analysis of the results of each step of the previous research activities	Confirming/refuting the applicability of the guidelines in a real-life application	RQ3

5 SecureDIS: A FRAMEWORK FOR SECURE DATA INTEGRATION

This research conjectures that considering SPT together in designing a DIS will reduce data leakage threats in these systems. Hence, this section presents a Secure Data Integration System (SecureDIS), an architectural framework that consists of several components, each of which includes a set of guidelines designed specifically to prevent data leakage. The following subsections discuss how SecureDIS was created and evaluated.

5.1 SecureDIS Construction

There are two parts to SecureDIS framework components and guidelines.

5.1.1 Constructing Components

Analysis of the previous studies (Gusmini and Leida 2011; Dicelie et al. 2001; Nachouki and Quafafou 2011) that build DIS with middle layer architecture, together with understanding the implications of using cloud computing, remote servers, and third parties to achieve integration contributed significantly to the formulation of SecureDIS components. The initial components of SecureDIS are data sources, the integration location, the integration approach and the data consumers, as adapted from the previous studies.

5.1.2 Creating SecureDIS Guidelines

The outcome of the careful review and analyses of the literature created an understanding of how several research areas relate to each other and contribute to securing a DIS. A DIS encompasses different levels: a low level that includes technical details of achieving secure and privacy-preserving integration to a much higher level of managing such a system to a medium level of engineering and designing a secure DIS. Synthesizing the results of the analysis, a set of guidelines were created. These guidelines are categorised into confidentiality, privacy and trust requirements and some of the guidelines overlap between these different categories. After constructing DIS components, the guidelines were grouped under each component. Each guideline was inspected individually and in relation to other guidelines in different components. This process of refining the guidelines was iterative, as many guidelines were moved around components and grouped differently until the final version was reached. A further analysis was conducted to link each guideline to the data leakage threats found in the data integration context and some guidelines were eliminated as they were out of the scope. The initial version of SecureDIS before evaluation can be found in (Akeel et al. 2013).

5.2 SecureDIS Evaluation

To confirm the SecureDIS framework and its guidelines, two expert reviews were conducted, one for the components and another for the guidelines. The following subsections provide an overview of the results of the confirmation.

5.2.1 Confirming SecureDIS Components

Based on the results of the first expert reviews of SecureDIS components, two additional components are proposed: *Security Policy* and *System Security Management.* Security policy was separated from other components of the system due to its importance in governing the security, privacy and trust of the DIS, which is supported by the work of Bhowmick, Gruenwald, Iwaihara, *et al.*(2006). System Security Management is added to ensure security measures are in place and to manage them, which is needed in any information system. Figure 1 shows SecureDIS framework after the first expert reviews.

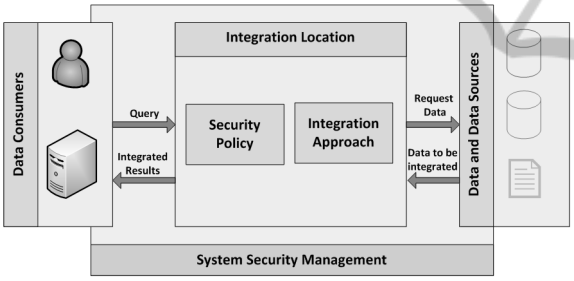

Figure 1: SecureDIS framework componenets.

5.2.2 Confirming SecureDIS Guidelines

To confirm and extend the proposed guidelines and answer the research questions, SecureDIS is planned to be reviewed by a second group of experts. The results will help in reshaping SecureDIS to an accepted version that can be useful to the system analysts and designers.

6 EXPECTED OUTCOME

The expected outcome of this study is a confirmed framework and set of guidelines, namely SecureDIS, which are comprehensive, practical and applicable to different contexts including cloud-computing environment. SecureDIS aims to help systems analysts and early designers in their analysis phase of building systems that considers security by design to prevent data leakage threats.

7 STAGE OF THE RESEARCH

The research questions RQ1 and RQ2 were answered. The current stage of the PhD is to customise and apply SecureDIS to a real-life context to assess its usefulness in preventing data leakage threats in DIS contexts. Possible case is an organisation using cloud services to integrate or store data coming from different resources. The results of the coming stage will help in answering RQ3.

REFERENCES

Akeel, F. et al. 2013. SecureDIS: a Framework for Secure Data Integration Systems. In: *The 8th International Conference for Internet Technology and Secured Transactions*. London, UK.

Akeel, F. Y. et al. 2014. Exposing Data Leakage in Data Integration Systems. *The 9th International Conference for Internet Technology and Secured Transactions (ICITST-2014)*, pp. 420–425.

Armellin, G. et al. 2010. Privacy preserving event driven integration for interoperating social and health systems. *Secure Data Management*, pp. 54–69.

Artz, D. and Gil, Y. 2007. A survey of trust in computer science and the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web* 5(2), pp. 58–71.

Avison, D. and Young, T. 2007. Time to rethink health care and ICT communications. *Communications of the ACM* (June 2007), pp. 69–74.

Barhamgi, M., Benslimane, D., Ghedira, C., Tbahriti, S.-E., et al. 2011. A Framework for Building Privacy-Conscious DaaS Service Mashups. In: *2011 IEEE International Conference on Web Services*. Washington DC, USA: IEEE, pp. 323–330.

Barhamgi, M., Benslimane, D., Ghedira, C. and Gancarski, A. L. 2011. Privacy-Preserving Data Mashup. In: *IEEE International Conference on Advanced Information Networking and Applications*. Biopolis, Singapore: IEEE, pp. 467–474.

Batty, M. et al. 2010. Data mash-ups and the future of mapping by. *JISC TechWatch*, pp. 1–45.

Begum, B. a. et al. 2010. Security policy integration and conflict reconciliation for data integration across data sharing services in ubiquitous computing environments. In: *International Conference on Computer and Communication Technology (ICCCT'10)*. Allahabad, India: IEEE, pp. 1–6.

Bhowmick, S. S. et al. 2006. PRIVATE-IYE: A Framework for Privacy Preserving Data Integration.

In: *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*. Washington, DC, USA: IEEE.

Boyens, C. et al. 2004. On privacy-preserving access to distributed heterogeneous healthcare information. In: *Proceedings of the 37th Hawaii International Conference on System Sciences*. Big Island, Hawaii, USA, pp. 1–10.

Van Den Braak, S. W. et al. 2012. Trusted third parties for secure and privacy-preserving data integration and sharing in the public sector. In: *Proceedings of the 13th Annual International Conference on Digital Government Research - dg.o '12*. College Park, MD, USA: ACM Press, pp. 135 –144.

Braghin, C. et al. 2003. Information leakage detection in boundary ambients. *Electronic Notes in Theoretical Computer Science* (78), pp. 123–143.

Carey, M. J. et al. 2012. Data Services. *Communications of the ACM* 55(6), pp. 86–97.

Clifton, C. et al. 2004. Privacy-preserving data integration and sharing. In: *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery - DMKD '04*. Paris, France: ACM Press, p. 19.

Cruz, I. et al. 2008. A Secure Mediator for Integrating Multiple Level Access Control Policies. *Knowledge-Based Intelligent Information and Engineering Systems*, pp. 354–362.

CWE 2013. CWE-200: Information Leak (Information Exposure). [Online] Available at: http://cwe.mitre.org/data/definitions/200.html [Accessed: 2 August 2013].

Dawson, S. et al. 2000. Providing security and interoperation of heterogeneous systems. *Distributed and Parallel Databases* (8), pp. 119–145.

Dicelie, J. J. et al. 2001. Data integration system.

Eze, B. et al. 2010. Policy-based Data Integration for e-Health Monitoring Processes in a B2B Environment: Experiences from Canada. *Journal of theoretical and applied electronic commerce research* 5(1), pp. 56–70.

Flechais, I. et al. 2013. In the balance in Saudi Arabia: security, privacy and trust. In: *Extended Abstracts on Human Factors in Computing Systems CHI '13*. Paris, France, pp. 823–828.

Fung, B. C. M. et al. 2012. Service-Oriented Architecture for High-Dimensional Private Data Mashup. *IEEE Transactions on Services Computing* 5(3), pp. 373–386.

Gollmann, D. 2006. *Computer Security*. Second Edi. John Wiley & Sons.

Gusmini, A. and Leida, M. 2011. A patent: Data Integration System.

Haddad, M. et al. 2012. Data Integration in Presence of Authorization Policies. In: *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*. Liverpool, UK: IEEE, pp. 92–99.

Halevy, A. et al. 2006. Data integration: the teenage years. In: *32nd International Conference on Very large data bases VLDB'06*. Seoul, Korea.

Harris, D. et al. 2007. Standards for secure data sharing across organizations. *Computer Standards & Interfaces* 29(1), pp. 86–96.

Hong, Y. et al. 2008. Protection of Patient's Privacy and Data Security in E-Health Services. In: *2008 International Conference on BioMedical Engineering and Informatics*. Sanya, China: IEEE, pp. 643–647.

Hu, Y. and Yang, J. 2011. A semantic privacy-preserving model for data sharing and integration. In: *Proceedings of the International Conference on Web Intelligence, Mining and Semantics - WIMS '11*. Sogndal, Norway: ACM Press.

Hung, P. 2005. Towards a privacy access control model for e-healthcare services. In: *Third Annual Conference on Privacy, Security and Trust*. Andrews, New Brunswick, Canada.

ISO 2014. ISO/IEC27000: Information technology — Security techniques — Information security management systems — Overview and vocabulary. *BSI Standards Publication*.

Jawad, M. et al. 2013. Supporting Data Privacy in P2P Systems. *Security and Privacy Preserving in Social Networks*, pp. 1–51.

Jurczyk, P. and Xiong, L. 2008. Towards privacy-preserving integration of distributed heterogeneous data. In: *Proceedings of the 2nd PhD workshop on Information and knowledge management*. Napa Valley, California, USA, pp. 65–72.

Khattak, Z. et al. 2012. Evaluation of Unified Security, Trust and Privacy Framework (UnifiedSTPF) for Federated Identity and Access Management (FIAM) Mode. *International Journal of Computer Applications* 54(6), pp. 12–19.

Lamb, P. et al. 2006. Role-based access control for data service integration. In: *Proceedings of the 3rd ACM workshop on Secure web services - SWS '06*. Alexandria, VA,USA: ACM Press, pp. 3–12.

Manan, J. A. et al. 2011. Security, Trust and Privacy–A New Direction for Pervasive Computing. In: *Proceedings of the 15th WSEAS international conference on Computers*. Stevens Point, Wisconsin, USA, pp. 56–60.

McGraw, G. 2004. Software security. *IEEE Security & Privacy Magazine*, pp. 80–83.

Meingast, M. et al. 2006. Security and privacy issues with health care information technology. In: *Proceedings of the 28th IEEE Annual International Conference of Engineering in Medicine and Biology Society*. New York, New York, USA, pp. 5453–5458.

Mohammed, N. et al. 2011. Anonymity meets game theory: secure data integration with malicious participants. *The VLDB Journal—The International Journal on Very Large Data Bases* 20(4), pp. 567–588.

Morton, A. and Sasse, M. 2012. Privacy is a process, not a PET: a theory for effective privacy practice. In: *Proceedings of the 2012 workshop on new security paradigms NSPW'12*. Bertinoro, Italy, pp. 87–104.

Nachouki, G. and Quafafou, M. 2011. MashUp web data sources and services based on semantic queries. *Information Systems* 36(2), pp. 151–173.

Pasierb, K. et al. 2011. Privacy-preserving data mining, sharing and publishing. *Journal of Medical Informatics & Technologies* 18, pp. 70–76.

Philip Coppel Qc 2012. The Data Protection Act 1998 & Personal Privacy. *Statute Law Society* 499(19 March 2012), pp. 1 – 31.

Pistoia, M. et al. 2007. When Role Models Have Flaws : Static Validation of Enterprise Security Policies Introduction : RBAC Systems. In: *29th International Conference on Software Engineering*. Minneapolis, MN, USA.

Pon, R. and Critchlow, T. 2005. Performance-oriented privacy-preserving data integration. *Data Integration in the Life Sciences*, pp. 240–256.

Prakash, V. and Darbari, M. 2012. A Review on Security Issues in Distributed Systems. *International Journal of Scientific & Engineering* 3(9), pp. 1–5.

Ray, S. S. et al. 2009. Combining multisource information through functional-annotation-based weighting: gene function prediction in yeast. *IEEE transactions on bio-medical engineering* 56(2), pp. 229–36.

Reeve, A. 2013. Cloud-Based Data Integration Adds Concerns about Latency and Security [Online] Available at: http://data-informed.com/cloud-based-data-integration-adds-concerns-about-latency-and-security/ [Accessed: 4 February 2014].

Ristenpart, T. et al. 2009. Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds. In: *Proceedings of the 16th ACM conference on Computer and communications security*. Chicago, Illinois, USA.

Ross, R. et al. 2014. Systems security Eegineering an integrated approach to building trustworthy resilient systems. *NIST Special Publication (800-160)*, p. 121.

Russom, P. 2008. Data Integration Architecture: What It Does, Where It's Going, and Why You Should Care [Online] Available at: http://tdwi.org/articles/2008/05/27/data-integration-architecture-what-it-does-where-its-going-and-why-you-should-care.aspx.

Saeed, M. Y. et al. 2014. Insight into Security Challenges for Cloud Databases and Data Protection Techniques for Building Trust in Cloud Computing. *Journal of Basic and Applied Scientific Research* 4(1), pp. 54–59.

Takabi, H. et al. 2010. Security and privacy challenges in cloud computing environments. *IEEE Security & Privacy Magazine* (December), pp. 24–31.

Tian, Y. et al. 2011. Dynamic content-based cloud data integration system with privacy and cost concern. In: *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference on - CEAS '11*. Perth, Western Australia, Australia: ACM Press, pp. 193–199.

Treglia, J. V. and Park, J.S. 2009. Towards trusted intelligence information sharing. In: *Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics - CSI-KDD '09*. Paris, France: ACM Press, pp. 45–52.

U.S. HHS 1996. Health Insurance Portability and Accountability Act (HIPAA) [Online] Available at: http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/.

Watson, D. 2007. Web Application Attacks. *Network Security* (October), pp. 10–14.

Whang, S. and Garcia-Molina, H. 2012. A model for quantifying information leakage. *Secure Data Management*.

Xiong, L. et al. 2007. Preserving data privacy in outsourcing data aggregation services. *ACM Transactions on Internet Technology* 7(3), p. 28.

Yau, S. and Chen, Z. 2008. Security policy integration and conflict reconciliation for collaborations among organizations in ubiquitous computing environments. *Ubiquitous Intelligence and Computing*, pp. 3–19.

Yau, S. S. and Yin, Y. 2008. A Privacy Preserving Repository for Data Integration across Data Sharing Services. *IEEE Transactions on Services Computing* 1(3), pp. 130–140.

Youssef, A. and Alageel, M. 2012. A Framework for Secure Cloud Computing. *International Journal of Computer Science* 9(4), pp. 487–500.

Zhang, D. Y. et al. 2011. Modeling and evaluating information leakage caused by inferences in supply chains. *Computers in Industry* 62(3), pp. 351–363.