Improving Database Security for Relational Database Management Systems

A dissertation submitted in partial fulfilment of the requirements for the degree of MSc Computing

Derek Colley

Student identifier: Y3322990

September 2018

Word count: 14,681

(excludes preliminary pages, abstracts, references and appendices)

Abstract

Relational Database Management Systems (RDBMSs) are software systems concerned with the management and administration of data. Data security can be defined as the continuance of confidentiality, integrity and availability in an RDBMS so that data remains protected from internal and external threats. In this research, we argue that data security is an underdeveloped sector of information security with unique challenges that stem from the construction of RDBMSs using different paradigms to other software applications, and that due to the growing value of data as a target for malicious actors, data security has become an issue of paramount importance.

Through a combination of primary research and literature review, we examine current attitudes and opinions in the data security field and seek information and clarity on forming best practices across the many different aspects of data security matters. We propose, design and populate the *Data Security Framework* (DSF), a flexible and comprehensive information model that encompasses the aspects of data security applicable to both industry practitioners and data researchers. The DSF is intended to serve as a central resource of information for both groups and establish a clear difference between data security matters and information security in general.

We also conduct a GAP analysis to discover areas in which academic research opportunities exist from industry developments driven by a high pace of technological change, and conversely where academic research has yielded potential opportunities to implement better data security controls through the novel application of new techniques. Finally, we discuss and present the technological and cultural themes that emerged from this research and speculate on potential future research directions that could augment and improve the outcomes.

Contents

List of Figures and Tables

Acknowledgements

I would like to acknowledge the invaluable assistance of my tutor, Dr. Chris Tucker, in providing continual feedback and useful guidance, both in the subject material and in preparing this dissertation. Thank you.

I would also like to thank my family for their understanding, patience and support during my studies.

Finally, I would like to thank all those who took part in contributing their time and knowledge by participating in the surveys, and without whom this research would have been impossible.

1. Introduction

1.1. Background to the problem/issue

Relational database management systems (RDBMSs) are software products designed to store and manage data (Astrahan et al., 1976; Codd, 1990; Coronel and Morris, 2016). By relational, this means collections of related data points which can be stored and queried by client applications with the interfaces supplied by the RDBMS using set-based operations in a format called Structured Query Language (SQL) (Lu et al., 1993).

As a persistent store of data, unauthorised access can lead to loss of revenue, reputation or produce other damaging effects for the owning institution (Afyouni, 2006) such as unauthorised data observation, modification and unavailability (Bertino and Sandhu, 2005), or loss of confidentiality, availability or integrity (Ge et al., 2005). Databases are routinely targeted for attack (see Section 2.2 for examples) due to the value of the payload should the database become compromised, and so ensuring adequate data security is both of academic and industrial interest.

Data security is essential to the wellbeing of any organisation. Repeated case studies and examples from industry spanning more than 10 years show the breadth and depth of the impacts – these are manifested particularly in reputational damage, leading to loss of custom, and both legal and financial liability. An early example is the loss of over 94,000,000 records through internal unauthorised disclosure by AOL, occasioned by an ex-employee who attempted to sell the data to a third party (Machanavajjhala and Reiter, 2012), undermining public confidence in AOL's ability to maintain data privacy and which could have contributed to their eventual decline. In 2013, Adobe, a prominent software house, confirmed over 38,000,000 usernames and encrypted passwords were stolen, alongside a significant proportion of the source code for its popular Photoshop product (British Broadcasting Corporation, 2013). The latter point is important – any competitor in possession of Adobe's intellectual property has an advantage when creating or cloning their software, to the original maker's detriment. More recently, the company Under Armour suffered a hack of over 150,000,000 usernames, passwords and e-mail addresses for the MyFitnessPal product, which measures users' calorific intakes and exercise regimens (Under Armour Performance Inc., 2018). This has a chilling implication; this information is an example of how biometrics can be stolen, allowing malicious actors access to the most intimate of health details for millions of users.

RDBMSs are the most common type of data repository in the world (Solid IT Consulting and Software Development gmbh, 2018) and follow standard essential design principles, but vary in their implementation. Due to this heterogeneity, security controls also vary between platforms. However, within the industry there are commonly accepted best practices in security management. These can stem from external standards such as ISO 27001 (International Standards Organisation, 2013) and NIST 800-53 (National Institute of Standards and Technology, 2013); held as tacit knowledge by practitioners; detailed in books (Basta and Zgola, 2011; (Gaetjen et al., 2015); studied in academic literature (Bertino and Sandhu, 2005; Jajodia, 1996; Pernul, 1994); or specified in the trade press (Microsoft Corporation, 2012). Some of these may not necessarily have been implemented in RDBMS systems. There are also more abstract concepts such as the C.I.A. principles (confidentiality, integrity, availability) (Afyouni, 2006; Olivier,

2002, Tchernykh et al., 2016) which are well understood in the industry (Microsoft Corporation, 2005, 2005b; Oracle Corporation, 2017) but for which differing controls are implemented.

Through interviews with industry practitioners, questionnaire engagement and literature analysis, this research shows that there is a disparity between how data security principles are applied by practitioners in industry compared to defined best practices in data security and the direction of academic research outcomes. By extension, this has led to a diaspora of security implementations that are not correlated with the literature and provides the motivation for developing what is termed the *Data Security Framework* (DSF), the key deliverable of this project. It is hoped that the DSF could a) benefit practitioners by providing a reference guide for database security implementation, and b) benefit academics by providing a range of new directions for further research.

1.2. Justification for the research

Developments in the RDBMS field were first led by academic experts such Codd (1969), Date (1986) and Stonebraker (1986), leading to extensive research output in relational database theory. However, new developments in the database arena are often led by software suppliers – cases in point include recent product features such as 'stretch' databases which can reside in cloud systems and on-premise simultaneously (Microsoft Corporation, 2016); machine learning integration augmenting hybrid RDBMS and application systems (Oracle Corporation, 2017); and developments in containerisation of RDBMS platforms (Docker Inc., 2017). Given that the collection and analysis of 'big data' is now a popular motivator in technological development, meaning in part the merger of relational and non-relational data sources in a fashion that considers the velocity and scale of the data, the importance of managing the security of such large and disparate estates makes this issue particularly timely.

Each innovation in data management brings new security challenges which may not be widely understood and which, in some cases, lack significant research pedigrees that verify or test the underlying assumptions to help understand and mitigate the security risks at all stages of implementation. New steps taken by industry result in the continual need for the reassessment of security considerations, but without an overarching reference framework for best practice, this is very difficult.

Databases are ubiquitous, and large database software suppliers like Oracle, Microsoft and Google are in a dominant market position to produce original research output. This argument by extension applies to data security research, a subset of general research. For example, a discussion on the impact of data credential theft is presented by individuals employed and funded by Google (Thomas et al., 2017), and a patent for controls built on original data security research is presented by Oracle Corporation (Kirti et al., 2017). The continuance of rapid technological change, led by the industry, appears to be a major contributing factor to the disparity between what is written about and understood to be best practice in data security, and actual understanding and implementation of the same by the customers, administrators and users of these systems.

From the practitioner's perspective, what makes good data security is a multi-layered question that can be answered in different ways depending on the context of the observer. From the perspective of the Data Protection Officer under GDPR (European Union, 2016), data security is defined solely by the requirements of the legislation, which includes, for example, the mandatory use of encryption. For the Database Administrator (DBA), good data security is in large part a subjective matter and may mean implementing technical controls such as role-based authentication, auditing and proactive monitoring to ensure confidentiality, integrity and availability. For the business owner, good data security may be viewed solely in terms of the risks presented to the business and the ways in which these risks are mitigated – these ways do not necessarily need to be technical in nature. Insurance, for example, as risk mitigation is suggested by Calder and Watkins (2005) and has no technological basis. Withdrawal of sensitive data from publicly-accessible systems to isolated internal platforms is also another design-based, rather than implementation-based, answer.

This is not to say that good database security standards do not exist. There are a range of competing standards; common in the UK is ISO 27001 (International Standards Organisation, 2013), which outlines IT security in the context of an Information Security Management System (ISMS). However, rather than mandating specific controls, the standard mandates that controls should be in place; albeit some controls are provided as guidelines in ISO 27002. This is a key difference and means that within an organisation, the fact that controls *exist* can be prioritised (wrongly) over the *effectiveness* of the controls themselves. Thus, ensuring an organisation's compliance to standards could take priority over assessing how adequately the standards are applied in practice. This point is also made in the literature. To quote Boehmer (2008), '*…registering an ISMS still says nothing about the quality and performance of its implementation*'. Or Sharma and Dash, 2012: '*…ISO 27001 is a management standard, not a security standard. It provides a framework for the management of security within an organization but does not provide a 'Gold Standard' for security, which, if implemented, ensures the security of an organization.*'.

This lack of detail on controls in the standard presents a problem for the practitioner looking for specific advice on how to ensure data security. Although data security controls to some extent are platform-dependent and so cannot be universally stipulated, it is left to the practitioner to research data security protections using the standards as guidelines, augmented by their own knowledge, experience, and pre-existing organisational processes. This could potentially result in a gap between an organisation's compliance status and the practical effectiveness of individual controls. As discussed in Section 2.2.2, insufficient data security protections, regardless of the organisation's compliance status, lead to high-impact incidents that damage enterprises. This favours the argument that there is a need for a more comprehensive set of universal data security controls and guidelines which are of everyday utility to the practitioner in the field.

1.4 Scope of the research

This research examines, through primary research, the extent to which different database security standards are executed through engagement with industry practitioners. There are some natural limitations:

- The scope is not intended to apply to all matters within IT security, which is a broad topic. Although database security is a subset of IT security, the nature of the technology and the assets (data) pertaining to databases mean that a different set of considerations apply. However, where applicable guidance from general IT security best practices has been incorporated into the research.

- This research is intended to cover security in relational databases rather than security for all (including non-relational) databases. The reason is that relational and non-relational databases have very different underlying principles and architectures.

- The Data Security Framework is the key deliverable of this research. This framework is defined fully further in this document. With 16 categories, 114 concepts and over 700 sections of information, it is not possible to present this data within this document in full, and neither has a full implementation been carried out. However, in lieu of this omission, two theoretical implementations and one partial practical implementation are provided which demonstrate the feasibility of a full implementation.

1.5 Outline of the dissertation

In ***Chapter 1***, we introduce relational databases and their place in industry is discussed. The primary security concerns are summarised, and we propose that there is a divergence in how database security is implemented in industry, driven by subjective interpretation of security standards, contending that current standards do not provide a sufficient level of detail to enable consistent cross-platform application of controls. We further propose that there exist gaps between the academic understanding of database security principles and the practitioner-led implementations of database security. We justify the research and introduce the Data Security Framework (DSF), a taxonomy of database security concepts codified and arranged into a form that allows its use as a reference guide for practitioners and as a source of information for academic researchers when considering future research directions.

In ***Chapter 2***, we expand further on the practical problem introduced in Chapter 1. We look at the threats and impacts to business from database security failures. In Section 2.2.3, we discuss existing information security models, from the early prototypes of the 1970s-1980s to more recent proposals including the GRC (2009), ISO 27001 (2013, 2016) and COBIT (2017), and critique their applicability to database security. We define the research aim and objectives, and create a roadmap of tasks and deliverables mapped back to the aim and objectives and presented both visually and narratively. Finally, the two key deliverables, the DSF and the GAP analysis, are defined.

In ***Chapter 3***, we describe the research methodology and justify the choices made. We chart a course through from philosophy through to techniques before describing how the research outcomes from each technique are used in conjunction. In Section 3.1, a table is provided mapping the philosophy, approach, research type, methodology, method, technique and objectives to one another, including method justification. Section 3.3 describes the research procedures in detail, including construction, validation and distribution of the research instruments, the approach taken to literature review and interview planning. Finally, the ethical considerations are discussed.

In ***Chapter 4***, we analyse the data collected. This chapter is split into the analysis of the questionnaire data, the interview data and the literature review data. Using a combination of statistical (empirical) analysis techniques, thematic analysis and grounded theory, we analyse the data. Finally, we present the structure and content of the DSF and GAP analysis deliverables.

In ***Chapter 5***, we draw conclusions on the effectiveness of the research. We present 3 example implementations of the Data Security Framework. We also present the conclusions from the GAP analysis in SWOT format detailing where specific technologies or database security techniques do not have extensive literature coverage, and conversely showing where existing database security literature contains ideas ripe for implementation, which leads to a discussion of further directions for future research and development. Finally, we discuss the implications of the research in terms of the applicability of the deliverables to practitioners and researchers, and finish with reflections on the research experience.

References are available in ***Chapter 6***, and the extended abstract is available in ***Chapter 7***.

All appendices follow thereafter.

1.6 Chapter Summary

In this chapter we introduced the subject area and provide context for the research to follow. We justified the research, showing it is topical, relevant and important in the field of information security. We outlined the scope of the research and provided an outline of the remainder of the dissertation.

2. Research definition

2.1. The practical problem

At present, there are a wide range of database security best practices being applied in organisations. These can range from minimal implementations using out-of-the-box security settings (Okman et al., 2011), to complex, detailed ISMSs compliant to industry standards. However, data breaches can and do occur regularly and have a negative effect on the ability of organisations to conduct business (Afyouni, 2006; Bertino and Sandhu, 2005).

The practical problem in brief is the inability of the practitioner to access a universal standard of data security, coupled with the inability of the database researcher to have access to the latest best practices within the industry. This problem is exacerbated by several factors. Agile software development, where the emphasis is on speedy releases and continuous integration, can suffer from lack of security governance (Rindell et al, 2015), although proponents argue that fast development cycles and responsiveness to change are more important than robust security testing (Abrahamsson, 2017). Other factors are insufficient control specifications in standards, as described in Section 1.2.

Another contributing factor is the variance between the perceived importance of security controls from the perspective of the practitioner. This variance is observed in the results of the survey, discussed in Section 4. This difference can be explained because of the lack of an overarching framework to govern the implementation of database security standards.

2.2. Existing relevant knowledge

2.2.1. General context

Information security (IS) is a topic that has evolved into a large research subject area in parallel to its development as a specialism within the information technology industry. However, there is a gap in the effective implementation of good IS practices. According to a report commissioned by the UK Government (Klahr et al., 2017), half of all businesses have not put in place basic security controls to mitigate cybersecurity risks. Additionally, 67% do not have a formal risk management process; 80% have not had any specific cybersecurity awareness training; and 87% do not require their suppliers to adhere to good practices in cybersecurity management.

Consequently, news of system breaches are common, and there is a lucrative market in stolen data (Romanosky et al. (2014) cite Privacy Rights Clearinghouse (2012)). Layton and Watters (2014) present estimated costs of USD$1,941,272.05 and USD$335,981.04 respectively for the Telstra data breach of 2011 (Telstra, 2011), caused by a website vulnerability leaving open access to customer records; and the LinkedIn data breach of 2012 (The Register,

2012) resulted in more than 6.4 million publicly-leaked credentials including hashed passwords from database records, aiding widespread phishing attacks.

The threat to data confidentiality, integrity and availability is getting bigger and more sophisticated. These are not limited to lone actors but can also be co-ordinated at the state level. Snegovaya (2015) defines the term 'reflexive control' to mean a method '*by which a controlling party can influence an opponent … by interfering with its perceptions*' and alleges it is a primary strategy of a major superpower. This strategy can include techniques such as leaking sensitive data to the public or disrupting systems on which an entity is highly dependent. Such techniques can be highly automated; the Linux Kali distribution (Offensive Security, 2018) is a publicly-available toolkit with hundreds of penetration test tools included. SQLMAP (Damele and Stampar, 2018) is a toolset compatible with 13 different RDBMS products that allows one to carry out highly-automated attacks. The high availability of these tools and the proliferation of the ease to carry out attacks increases the threat level to organisations.

Arlitsch and Edelman (2014) discuss the implications of issues such as commercial data loss and assert that security breaches are often attributed to the exploitation of improperly managed systems to gain sensitive data. With similar views, Custer (2010) echoes the view that adverse incidents continue to rise, citing various public breaches as evidence for this view.

2.2.2. Database security

Payloads from data breaches which have attracted public interest often include data sourced from an internal database. The Ashley Madison data breach in 2015 (Mansfield-Devine, 2015), the Sony Playstation Network data breach in 2011 (Cachin and Schunter, 2011) and the Equifax breach of 2017 (Gressin and Charleston, 2017; Woo, 2017) demonstrated how once the attackers had broken the system perimeter they were able to access database information without significant impediment.

Database platforms (RDBMSs) provide the means to ensure high-grade security. For example, Microsoft SQL Server meets the FIPS 140-2 encryption standards (Microsoft Corporation, 2016b) and the Common Criteria Certification Evaluation Level 4+ (Common Criteria, 2018), as does Oracle Database (Oracle Corporation, 2018) and all major platforms support encryption, auditing, user access controls, change tracking and more. However, it is up to the owners of these systems how the data is operationally managed, and misconfiguration or the omission of controls for data security is the responsibility of the owning party. It is their implementation that varies, rather than the facilities available in the RDBMS, due to the absence of a detailed universal framework.

Academic research is continually contributing new ideas in the data security space. For example, Bamrara (2015) addresses potential controls to augment current security features, such as security categorisations in the style of Bell-LaPadula (1976). The quantitative methodology posited in Yasnoff (2016) is an attempt at categorising the severity of the risk of database breaches, in contrast to the Layton and Watters (2014) cost-based approach. There appear to be a variety of proposals to bolster database security. Vavilis et al. (2015) present an anomaly detection process to enable both rule- and behaviour-based detection of potentially suspicious events through the analysis of SQL

queries. Gupta et al. (2016) present a method of authenticating a database user by having the user select a predetermined image from a set which is possibly useful as a technique in combination with other more secure methods.

There is overlap in both practitioner and academic literature around best practices for database security. One important paradigm is the principle of least privilege (Saltzer and Schroeder, 1975), controlled using techniques such as role-based access (Shete and Kulakrni, 2015). This work was built upon by others including Schneider (2003) who proposed a 'reference monitor' process to enforce fine-grained access control on program extensions, beyond the operating system. This translates to RDBMS systems, which have a separate secondary layer of security which can be implemented by OS inheritance; by separate, basic authentication (username and password); or through more obtuse techniques such as multi-stage authentication.

Databases also suffer from some attack vectors peculiar to RDBMSs, such as SQL injection (Kindy and Pathan, 2011), CSV injection (Kettle and Context Information Security, 2014) and homoglyphic attacks (Roshanbin and Miller, 2011; Qiu et al, 2010). Another issue is the disconnection between the application and database layer, which can be mitigated using techniques such as application roles (Beauchemin, 2012), row- or cell-level security (Rask et al, 2005), and strategies such as the separation of duties between different roles (Gick and Richins, 2008).

2.2.3.Existing IS models

The Biba model (Mitre Corporation, 1975) is an early structure that defined the reference monitor (the process that decides access levels), access domains and multi-level access control. From this idea came lattice security and state machine representation of security (summarised in Landwehr, 1981), followed by Clark et al (1991) who discuss granularity of object permissions which can be split into mandatory (MAC) and discretionary (DAC) types. Ge et al (2004) cite Clark-Wilson (1986) and the eponymous security model, discussing the applicability to RDBMS systems. However, these models have become dated and, although appropriate to the software systems of their time, suffer from flaws in the context of modern computing. One such flaw is a focus on the assignation of correct permissions once the user is inside the system boundary at the expense of consideration given to external threat actors that are able to access these systems externally, enabled to do so in modern times through Internet connectivity.

More recently, Hill (2009) proposes the Governance, Risk Management and Compliance (GRC) framework which has a focus on data protection – guidance includes data classification, implementation of archival methods, and introduces *responsiveness* to the C.I.A. triangle. This framework separates data from information and emphasises the importance of security at multiple levels – business, process and technical. Some of the technical guidance is dated but stands as a rigorous treatment of the subject overall, however responsiveness is arguably synonymous with availability, since availability is measured in several ways within industrial systems, for example by Service Level Agreements (SLAs), which may dictate specific availability targets for a given system. These targets will include uptime percentage, capacity for concurrent connections or queries, and potentially response time measured in the order of milliseconds (particularly in the context of high-availability websites).

Data security models have parallels with wider sets of IS principles, mandated in various standards, including ISO 27001 (International Standards Organisation, 2013 and 2016), COBIT (ISACA, 2017), PCI-DSS (PCI Security Standards Council, 2017) (for financial transactions) and NIST 800-53 (NIST, 2013). Correspondingly, there have been supplementary texts from authors knowledgeable about their implementation. Calder and Watkins (2005) expound upon the ISO 27001 standard with examples but stop short of adding a full architecture of controls. Stallings and Brown (2012), in material for the CISSP qualification, focus more heavily on practical controls in ISMSs, detailing several attack techniques in the database layer (e.g. output leakage) and provide a range of interesting directions when considering their application to relational databases.

2.3. Aim, objectives, methods, tasks and deliverables

2.3.1. Research Aim

To investigate potential disparity between implementation and best practice (in the contexts of both academic research and industrial standards) in data security management; to construct a reference framework of database security controls for use by both the academic research and industrial professional communities; and to identify gaps in research or practice which could contribute to stronger database security measures, the whole to culminate in the production of a Data Security Framework to aid future data management researchers and practitioners.

2.3.2. Research Objectives

In support of the research aim, the following 5 objectives have been defined:

O1. Conduct a literature review synthesising and summarising the progress made in database security research;

O2. Conduct a comprehensive search of the practitioner literature and conduct primary research by way of semi-structured interviews and a survey to determine a collection of best practices in database security;

O3. Consolidate the findings of 1 and 2 to produce a Data Security Framework which will be a detailed taxonomy, or catalogue, of database security concepts, best practices, methods and techniques together with their linkages, contextual information and examples;

O4. Produce a GAP analysis between the academic and industrial states of database security and propose a catalogue of findings for future research or implementation;

O5. Reflect on wider factors such as cultural or technological trends, research and implementation challenges to help refine and validate the model, proposing future work.

2.3.3. Tasks

The research methods selected can be divided into primary research, surveys using the techniques of questionnaire and interviews; and secondary research, the ongoing literature review.

Figure 2.1 shows how these activities inter-relate to produce the deliverables. Each deliverable is mapped to one or more objectives, described in Section 3.1.

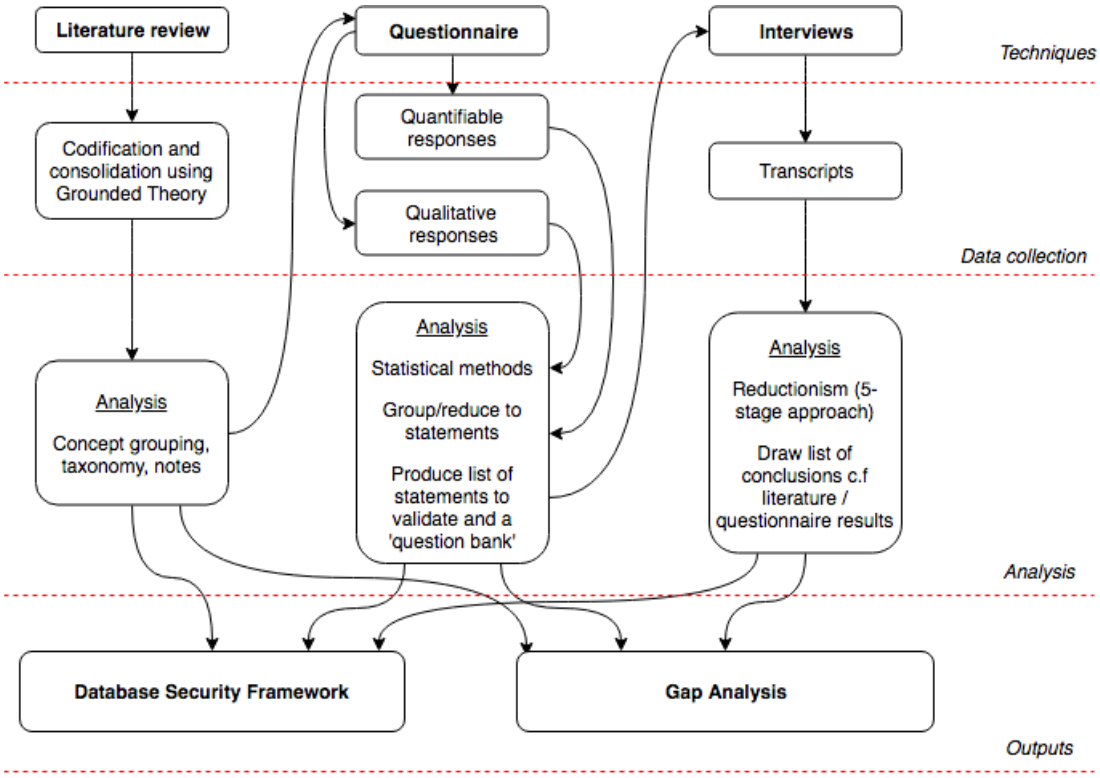


Figure 2.1: From techniques to deliverables

2.3.4. Deliverables

The key deliverables from this research are described in objectives O3 and O4. These are a new Data Security Framework, termed DSF, and the production of a GAP analysis of data security topics between the practitioner and academic spheres. These are described below.

2.3.4.1 The Data Security Framework (DSF)

The purpose of the DSF is to provide a comprehensive set of database security topics, best practices and current strategies, drawn together from both practitioner-led best practices and academic research sources. This means that researchers will be able to better identify (or exclude) underdeveloped areas of relational database research subjects, and practitioners will have a reference source that enables them to fully consider all aspects of database security when designing or maintaining systems.

2.3.4.2 GAP analysis

GAP analyses can be defined as summations of the current versus the desired state of affairs; in the context of this research, this is the gap between practitioner-led best practices and academic-led research, in both directions. In this context GAP stands for Good, Average and Poor, and indicates how well each topic is represented in each of the domains, industry and academia. This is an approach noted for its use in the PRINCE2 methodology (Bentley, 2012).

Use of the SWOT analysis technique (Strengths, Weaknesses, Opportunities and Threats), as pioneered by Learned (1969), is used in conjunction with GAP allocation to assess the overall landscape for each category in the DSF, by assessing the strengths, weaknesses, opportunities and threats of implementing (or conversely, researching) each topic and its contents.

The GAP analysis is presented in full in Section 4.3.3 and the SWOT analysis in Section 5.1.4.

2.4 Chapter Summary

In this chapter we have introduced data security and examined the problem of applying controls where current guidance is fractured and against a changing backdrop of threats. We specified the research aim, methods and deliverables, and the research tasks which will lead to their fulfilment.

3. Methodology

3.1. Methods and techniques selected

The pragmatic research philosophy was used as this is a good fit with the nature of the research, which is balanced more in favour of practical implementation-based features than abstract concepts. From this, an inductive approach was chosen as this aligned with the objectives which concern information discovery and synthesis, in contrast with the deductive approach (inferring specific conclusions from general reasoning) or the abductive, which focuses on drawing best guesses from incomplete data.

Consequently, in keeping with the pragmatic philosophy and inductive approach, the mixed methods research type was selected as this allows for both empirical and non-empirical methodologies to be used. Figure 3.1, an adaption of the 'research onion' (Saunders, 2009), shows this pathway between philosophy and methodology:

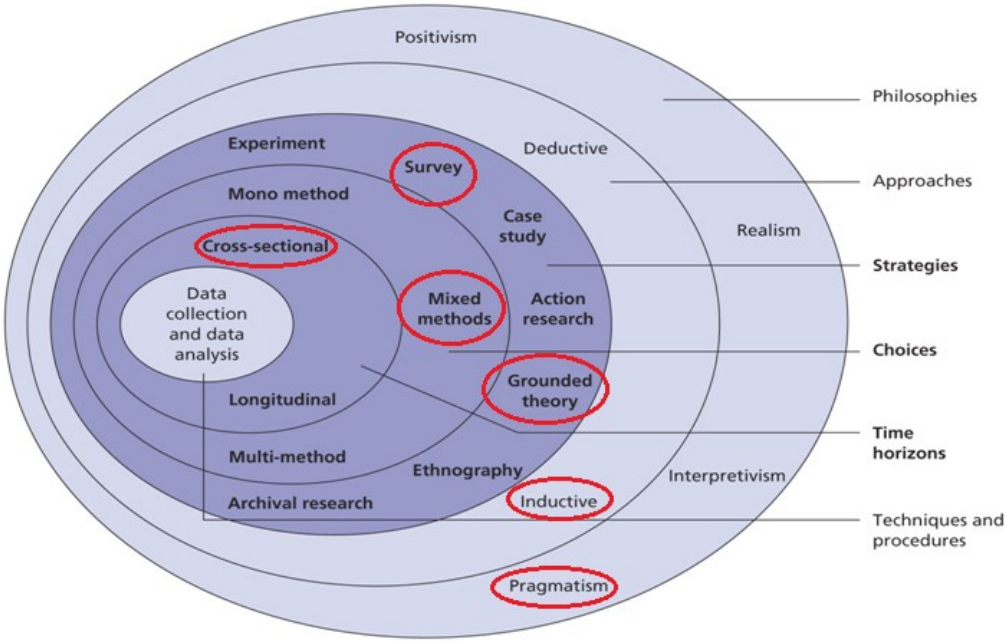


Figure 3.1: Adaption of the 'research onion' with the chosen project philosophy and methods shown.

The survey methodology, using a comprehensive questionnaire (administered electronically) and semi-structured interviews (face-to-face and video conference), aimed at a cross-section of database professionals and specialist academics was combined with non-empirical review, using re-interpretation and synthesis from existing literature to draw general conclusions from specific data, forming the foundation of the DSF. For the re-interpretive analysis and literature review, an adaption of Grounded Theory (GT) is used - an inductive method to explore, codify and synthesise volumes of non-structured information. This is used as the literature comprises of many heterogeneous sources in different formats covering different areas and with different foci, which GT is equipped to analyse.

The outcomes were explored with thematic analysis (Boyatzis, 1998) to triangulate the outcomes (Cresswell and Plano-Clark, 2007), so that areas of interest resulting from one research method were explored further using another. An example of this is that the outputs of the semi-structured interview provided new directions for literature exploration. The secondary research (literature review) helped in informing the structure of the survey techniques and was integral to the review stage, culminating in the delivery of the DSF and the GAP analysis between theory and practice as detailed in the Objectives.

The table below shows how the philosophy, approach, research type, methodologies and methods combine to produce the research tasks, and how they map to the Objectives.

Table 3.2: Mapping philosophy to research type, technique and objectives

Philosophy	Approach	Research type	Methodology	Method	Technique	Objective(s)
Pragmatism	Induction	Empirical (quantitative)	Survey	Questionnaire	Electronic administration	O2, O4, O5
		Empirical (qualitative)		Semi-structured interview	Face-to-face or video conference	O2, O4, O5
		Non-empirical	Review	Reinterpretivism	Adaption of Grounded Theory	O1, O2, O3, O5
				Literature review (secondary)	General reading	O1, O2, O4

Each method or technique in the table is described in more detail in the subsections below.

The survey methodology was of an explanatory type i.e. looking to confirm and account for the disparity between the industrial and academic sectors in database security practices. This means asking about probable causes for phenomena, as described in de Vaus (2002).

3.2. Justification of Research Methods

3.2.1. Survey – Questionnaire

Objective 2 includes the need to discover information on this topic directly from professionals in the field and academic specialists. In keeping with the inductive approach, the purpose of this method was to identify further research directions from general questioning, with a good sample size to preclude skew or incompleteness, to triangulate with another research method (interviews) on specific findings.

In criticism of this method, the questionnaire may be useful for surveying broad opinion but having a rigorous, measurable structure implies a finite set of pre-selected answers, and individuality of expression may be lost. There was also the risk that a modest sample size, or a sample consisting of an unrepresentative cross-section of participants, may damage the validity of the findings, which was observed in the results. There is an element of randomness in distribution and response that may have heightened this risk, mitigated in part by the inclusion of two qualifying questions.

3.2.2. Survey – Interviews

Data security is a wide-ranging topic with many principles appearing to be inherited from wider IT security principles and standards. This variety of different sources includes tacit knowledge amassed from experience, and the interview is used both for the triangulation upon specific topics and for the ability in a semi-structured interview to collect much more detailed information on these topics using a qualitative approach of guided conversation than a stricter quantitative approach. The interview method was selected as it is thought to be the most comprehensive way to gather otherwise-inaccessible data from willing participants in keeping with Objective 2, and to verify the findings from literature (Objective 1). The relative advantages of different types of interview are expanded further in the context of software engineering by Hove and Anda (2005) and Gubrium and Holstein (2017).

In criticism, this method is highly subjective and there are multiple risks. One risk includes the bias that may be injected into the interview unwittingly by the interviewer, if the interviewer shares some knowledge of the area of questioning – for example by leading answers, guiding the respondent in certain directions, or priming the respondent to answer in certain ways.

3.2.3. Literature Review, Synthesis and Re-interpretive Consolidation using Grounded Theory

Seeking to synthesise literature on data security requires a robust method of enquiry that allows the rapid assimilation of many sources of data and consequent reduction into a series of codified concepts that can be sorted and linked. Grounded theory (GT) is a research method expounded by Glaser and Strauss (1967) which, although originally aimed at data collection in social science, fits well for investigating large, disparate data sources and condensing these as described.

GT is not used in full. Rather, only the concepts of organising information by hierarchy, labelling and structuring concepts into trees, and writing 'memos' – *ad lib* commentary on the data – is used. A process workflow has been constructed (see Section 3.3.3) to enable this, which was modestly successful.

Criticisms of this approach are that it is was not originally designed for use in analysing academic literature; that it has proved quite time-consuming when organising data topics; and that in this specific context, each investigation into a topic opens an array of sub-topics for investigation, leading to scope creep and a certain amount of subjective

selection required in the topics to be followed. In practice, this GT approach was supplemented with a list of parent categories to help provide some structure.

3.2.4. Contrasting Methods

Consideration was given to other methods of enquiry. Using a case study could have worked well as an in-depth, industry-specific view of data management may have resulted in greater incorporation of the business elements of data security; likewise using an *in-situ* context rather than abstract research may have been conducive to obtaining a full picture of all factors in data security design. However, a suitable organisation was not available, and the case study would have excluded elements of data security not practiced by the organisation.

Another avenue explored but rejected was empirical, or laboratory-based, experimentation using RDBMS platforms to categorise the features present in each system, compare the findings, and formulate the DSF through the observation of the various implementations. However, this is a deductive approach, examining the generalities of each system to form the specific theory, rather than the inductive approach, examining the specific theories to form the generalities, and would consequently fail to include the insights offered by the extensive academic and trade literature available on information security standards.

3.3. Research procedures

3.3.1 Questionnaire construction, distribution and analysis

The questions posed in the questionnaire resulted from the topics discussed in the literature review and are closely related to the objectives. As it was planned that the questionnaire results would be analysed quantitatively, it was constructed in such a way as to force measurable answers using the Likert scale in most cases, with open questions interspersed to collect more detail.

The questionnaire was initially piloted to a pool of 5 selected individuals. Feedback was gathered on why certain choices were made and the implementation construction through email conversation with the participants. This helped to identify issues resulting in rewriting some of the questions to be less ambiguous; fixing formatting problems; fixing logic flow problems; and recasting some questions (e.g. the multiple-row questions on the Likert scales) as matrices.

The questionnaire was then distributed to a wider audience. Given that the questionnaire is targeted at data professionals, the deployment was made via a combination of posts on the professional networking platform LinkedIn, within 8 separate data-related specialist subject groups with an approximate estimated membership of 25,000 individuals and organisations. Additionally, a general post was made to the author's connections, a substantial portion of which are data professionals, totalling a potential pool of approximately 1,100 individuals, and to the Slack platform in a channel dedicated to data professionals.

Figure 3.3 illustrates the question flow which is designed to be completed by practitioners or academics. There are two principal flows depending on the occupation of the participant. Control flow logic is embedded to direct participants to certain questions based on their answers to other questions.

The colour key indicates which questions have outcomes that can be quantitatively measured and which that can be qualitatively measured. Each question is categorised, and each category can be mapped to helping to complete one or more objectives. The category to objective mappings can be found in Section 4.2.1, Figure 4.14.

The exact question wording is available as Appendix A together with selected screenshots of the SurveyMonkey implementation and the URL, https://www.surveymonkey.co.uk/r/2DZNG5G. All logic as described in the diagram has been encoded in the implementation.

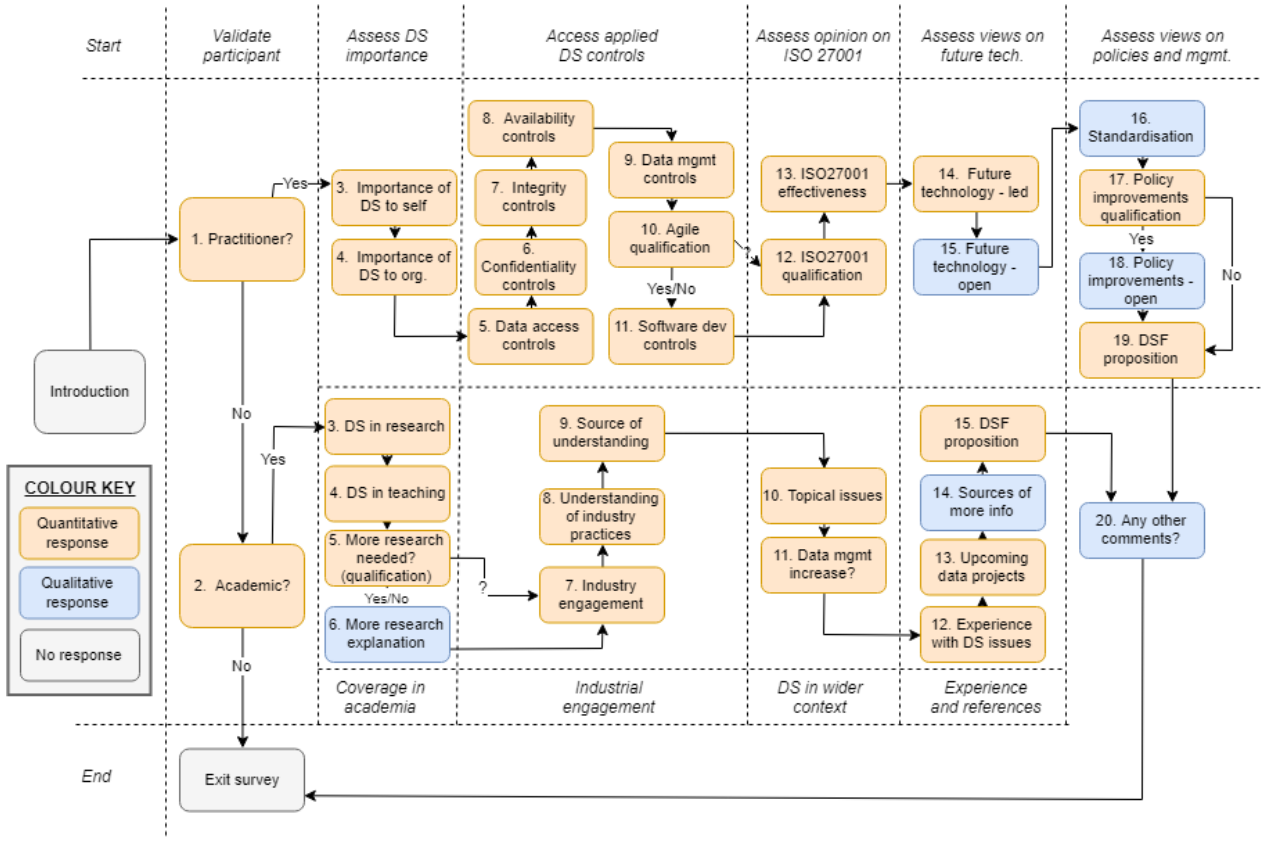


Figure 3.3: Flowchart describing the construction and process flow of the questionnaire.

The outcomes of the questionnaire are discussed in Section 4 using a statistical analysis of findings from quantitatively-measured questions; reductionist summarisation of qualitative answers; and conclusions drawn with reference to the other research method outcomes.

3.3.2 Interview planning

Three semi-structured interviews were planned and carried out successfully. These were focused on data security matters and began with a general discussion on how important database security is to the individual and their organisation, followed by an in-depth discussion on the techniques currently used by the individual. In general, the conversation moved on to the merits (or otherwise) of various techniques before diverging into related topics. The process and outcome of the interviews is described in Section 4.2.1. The questions were based on the plan presented in Appendix A.

The individual for the first interview was approached as a senior database administrator for a mid-size organisation with approximately 1,100 staff, responsible for maintaining a sizeable estate of database services. The interview was conducted via Skype. The second interviewee had a background as a data architect and was known only by association to a prior acquaintance of the author, which helped remove bias. The first two interviews were conducted by telephone. The third and final interview was with an academic lecturer with a background in data science. This interview was conducted via Skype. The final interviewee was not personally known to the author and responded to an invitation to participate sent through an academic networking facility.

The interviews were analysed using a simplification of thematic analysis as defined by Boyatzis (1998). The steps undertaken can be summarised as (based on Boyatzis):

1) Reducing the raw information – Paraphrase/summarise each unit of information
2) Identifying themes – Determine similarities between the units of information
3) Comparing themes – For when data from multiple interviews is available.

   Compare the themes from Step 2 across samples (interviews).
4) Create a code – Codify each theme e.g. by category or unique identifier
5) Integrate coded themes into other research findings

The Boyatzis approach was chosen as thematic analysis appears often in the literature of qualitative data analysis, and the author is a seminal authority in the field. The analysis is presented in Section 4.2.

The interviews were recorded (with informed consent). Boyatzis (1998) suggests that 60 minutes of interview can result in up to 20-40 pages of transcript. Given the time-intensive nature of this process and the limitations of the project, instead of transcription, the 5-step process was directly applied to the audio rather than transcribing, and the findings of this are presented in Section 4.1.

3.3.3 Grounded theory and the literature review

Although the research is dependent on the outcomes of the primary research methods (questionnaires and interviews), it is also heavily dependent on analysing the outcomes of the secondary research method, the literature review. For this reason, Grounded Theory (Glaser and Strauss, 1967) is a suitable approach to structuring the terms, categorising and linking notes to create several annotated hierarchies of topics and establishing through exploration a

clear understanding of the database security subject matter. This structured approach should also help ensure delivery of taxonomically-structured research results.

A process workflow was constructed (see Figure 3.4) to enable this, which was modestly successful.

The following diagram illustrates how data has been collected through literature review.

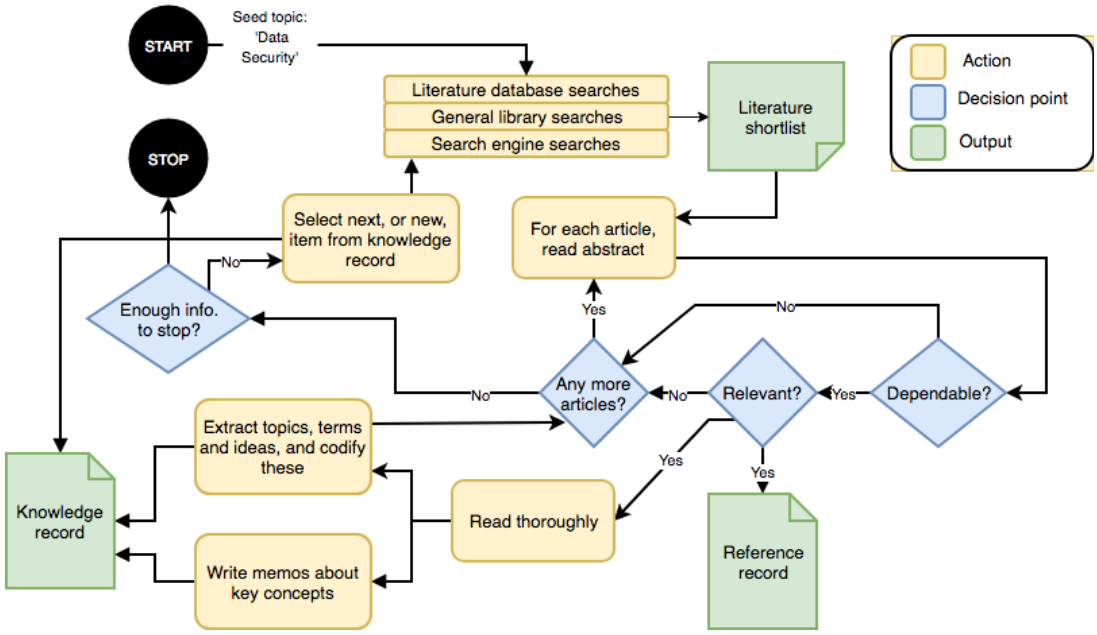


Figure 3.4: Flowchart illustrating method of secondary data collection.

The stop condition is a subjective judgment on whether enough information on a sub-topic has been collected to arrange that information into a meaningful set of data for entry into the DSF. This has the disadvantage of being non-repeatable (or non-deterministic) but adds the advantage of human decision-making and intuition to an otherwise algorithmic process.

In conducting the review, one discovery has been that maintaining a tightly arranged set of *précis* is easier in theory than in practice, since simultaneous work updating the current literature review and maintaining notes from other sources leads to difficulty when synthesising topics and drawing conclusions.

3.4 Ethical considerations

The research procedures that are subject to ethical consideration are the survey techniques – questionnaires and interviews.

For the questionnaire, the following points were stated at the beginning to provide reassurance over data governance to the participants:

Please note that no personal details are collected or retained as part of this survey, unless you choose to provide your email address at the end.

For any questions relating to this survey or on how the results will be processed, please contact Derek Colley at derek@derekcolley.co.uk.

For the interviews, the following points were explained to the participants before each interview began:

- They had a right to withdraw from the interview at any stage
- The interview was being recorded (and their specific consent for this was captured in the recording)
- The interview audio would be retained only as long as necessary for analysis and then be destroyed
- No personal data (name, employer, specific occupation) would be retained
- Their contributions would be anonymised and aggregated where possible in the research output

After interview analysis was completed, the audio files were deleted as planned, and no personal details have been stored. No participants chose to withdraw.

Other than the precautions detailed above over data governance, anonymity and right to withdraw, there are no further ethical considerations relevant to this research.

3.5 Chapter Summary

In this chapter we selected and justified the research methods in relation to the aim and objectives. We detailed the techniques used to collect primary research data and showed how the literature review was conducted. We also specified the ethical considerations of the research.

4. Analysis and Interpretation

4.1 Data Collection

4.1.1 Primary Research - Questionnaire

Data Collection and Cleansing

The design of the questionnaire was discussed in Section 3.2.1 and consists of 2 question tracks, one for industry respondents and one for academic respondents, which have 20 questions and 16 questions to answer respectively. These questions are, in the most part, standard Likert-scale questions (unforced) with the addition of a 'Don't know' option, arranged into matrices. The full question bank is available in Appendix A.

The questionnaire was deployed and gained 18 responses. The proportion of respondents who completed the entire questionnaire was 33%, with respondents spending a (mean) average time of 5 minutes and 30 seconds before exiting. Based on the pool of potential respondents, this is a smaller-than-expected response – this may be due to the delivery mechanism, the wording of the survey invitation, the time of day (or weekday) that the survey was deployed, or other factors. Two additional messages were sent encouraging respondents to complete the survey with zero additional replies.

Data from the questionnaire was downloaded as raw comma-separated values from the SurveyMonkey website, where it was imported into Excel and manipulated into a usable form. This included setting numerical data types on columns; truncating text where necessary; moving columns and rows to form a clearer view of the data; generating graphs and altering graph parameters to display the data; and calculation of aggregate values such as percentages from the data.

The results of questions 4 through 8, dealing with how well the respondents feel data management, confidentiality, integrity, availability and software development controls are implemented at their organisation are presented below. Analysis follows in Section 4.2.

Q4. Thinking about data access control, please rate how well or otherwise you feel the following practices are applied within your organisation.

	Values					
	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total
Password policy enforcement	1	1	7	5	0	14
Use of domain, rather than basic authentication, accounts	1	4	7	2	0	14
Principle of least privilege	1	4	8	1	0	14
Restriction of sysadmin / sa / privileged accounts	2	2	6	4	0	14
Auditing failed login attempts	2	2	4	4	2	14
Information classification	3	3	3	4	1	14
Formal authorisation procedures in place for access requests	3	2	6	3	0	14
Retrospective review of user accounts or access levels	4	4	4	1	1	14
Control of third-party access to organisational data	0	1	5	7	1	14
					Answered	14
					Skipped	4

	Percentages					
	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total
Password policy enforcement	7.1	7.1	50.0	35.7	0.0	100.0
Use of domain, rather than basic authentication, accounts	7.1	28.6	50.0	14.3	0.0	100.0
Principle of least privilege	7.1	28.6	57.1	7.1	0.0	100.0
Restriction of sysadmin / sa / privileged accounts	14.3	14.3	42.9	28.6	0.0	100.0
Auditing failed login attempts	14.3	14.3	28.6	28.6	14.3	100.0
Information classification	21.4	21.4	21.4	28.6	7.1	100.0
Formal authorisation procedures in place for access requests	21.4	14.3	42.9	21.4	0.0	100.0
Retrospective review of user accounts or access levels	28.6	28.6	28.6	7.1	7.1	100.0
Control of third-party access to organisational data	0.0	7.1	35.7	50.0	7.1	100.0
					Answered	77.8
					Skipped	22.2

Figure 4.1: Results from Question 4

Q5. Thinking about measures to preserve confidentiality, please rate how well or otherwise you feel the following practices are applied within your organisation.

	Values					
	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total
Use of encryption of data at rest (e.g. Transparent Data Encryption,	3	6	0	2	0	11
Use of encryption of data in flight (e.g. SSL channels, SFTP,	0	2	8	1	0	11
Granular security control (e.g. row-level encryption, table/view-	5	5	1	0	0	11
Use of intrusion detection systems at the database (or	2	3	0	4	2	11
Physical security (e.g. door controls in the server room; data	0	1	0	10	0	11
Security of backups	0	1	3	6	1	11
Regular software patching	1	2	5	3	0	11
Separation of test and production data	0	1	7	3	0	11
					Answered	11
					Skipped	7

	Percentages					
	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total
Use of encryption of data at rest (e.g. Transparent Data Encryption,	27.3	54.5	0.0	18.2	0.0	100.0
Use of encryption of data in flight (e.g. SSL channels, SFTP,	0.0	18.2	72.7	9.1	0.0	100.0
Granular security control (e.g. row-level encryption, table/view-	45.5	45.5	9.1	0.0	0.0	100.0
Use of intrusion detection systems at the database (or	18.2	27.3	0.0	36.4	18.2	100.0
Physical security (e.g. door controls in the server room; data	0.0	9.1	0.0	90.9	0.0	100.0
Security of backups	0.0	9.1	27.3	54.5	9.1	100.0
Regular software patching	9.1	18.2	45.5	27.3	0.0	100.0
Separation of test and production data	0.0	9.1	63.6	27.3	0.0	100.0
					Answered	63.6
					Skipped	36.4

Figure 4.2: Results from Question 5

Q6. Thinking about measures to preserve the integrity of the data, please rate how well or otherwise you feel the following practices are applied within your organisation.

	Values					
	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total
Regular checks for corrupted databases	3	1	3	4	0	11
Auditing of changes to business-critical data	2	6	2	1	0	11
Error or application logging	1	2	4	4	0	11
Error or application log reviews	2	3	3	2	1	11
Regular backups	0	0	2	9	0	11
Backup testing	1	3	4	2	1	11
Disaster recovery planning	2	2	4	2	1	11
Operational recovery planning (day-to-day issues)	1	2	5	2	1	11
					Answered	11
					Skipped	7

	Percentages					
	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total
Regular checks for corrupted databases	27.3	9.1	27.3	36.4	0.0	100.0
Auditing of changes to business-critical data	18.2	54.5	18.2	9.1	0.0	100.0
Error or application logging	9.1	18.2	36.4	36.4	0.0	100.0
Error or application log reviews	18.2	27.3	27.3	18.2	9.1	100.0
Regular backups	0.0	0.0	18.2	81.8	0.0	100.0
Backup testing	9.1	27.3	36.4	18.2	9.1	100.0
Disaster recovery planning	18.2	18.2	36.4	18.2	9.1	100.0
Operational recovery planning (day-to-day issues)	9.1	18.2	45.5	18.2	9.1	100.0
					Answered	63.6
					Skipped	36.4

Figure 4.3: Results from Question 6

Q7. Thinking about measures to preserve availability, please rate how well or otherwise you feel the following practices are applied within your organisation.

	Values						Percentages					
	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total
Database performance tuning	0	4	4	1	0	9	0.0	44.4	44.4	11.1	0.0	100.0
Database query performance tuning	0	5	2	1	1	9	0.0	55.6	22.2	11.1	11.1	100.0
Server-level redundancy (e.g. clustering)	1	4	3	1	0	9	11.1	44.4	33.3	11.1	0.0	100.0
Database-level redundancy (e.g. replication, mirroring)	1	3	3	0	2	9	11.1	33.3	33.3	0.0	22.2	100.0
Server-level monitoring	0	4	0	5	0	9	0.0	44.4	0.0	55.6	0.0	100.0
Database-level monitoring	1	5	1	2	0	9	11.1	55.6	11.1	22.2	0.0	100.0
Automatic failover mechanisms	1	5	2	1	0	9	11.1	55.6	22.2	11.1	0.0	100.0

Figure 4.4: Results from Question 7

Q8. Thinking about overall management of data within your organisation, please rate how well or otherwise you feel the following practices are applied.

	Values						Percentages					
	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total
Existence of a formal Information Security Management System	3	0	2	1	2	8	37.5	0.0	25.0	12.5	25.0	100.0
Regular review of security documentation	3	0	3	1	1	8	37.5	0.0	37.5	12.5	12.5	100.0
Security / audit reports regularly surfaced to management	3	0	3	1	1	8	37.5	0.0	37.5	12.5	12.5	100.0
Security incident management response procedures	3	0	1	2	2	8	37.5	0.0	12.5	25.0	25.0	100.0
User training on data security	1	2	2	3	0	8	12.5	25.0	25.0	37.5	0.0	100.0
Policy or procedures on third-party (e.g. supplier) data access	2	1	2	2	1	8	25.0	12.5	25.0	25.0	12.5	100.0
Asset tracking procedures	2	1	3	0	2	8	25.0	12.5	37.5	0.0	25.0	100.0
Mobile / own-device (BYOD) policies on data access	1	0	5	1	1	8						
Data sharing policies for staff (e.g. social media guidance)	1	2	1	4	0	8	12.5	25.0	12.5	50.0	0.0	100.0
					Answered	8					Answered	44.4
					Skipped	10					Skipped	55.6

Figure 4.5: Results from Question 8

Question 9 was a qualifying question asking whether respondents' organisations used Agile methodologies. Of 8 respondents, 2 answered 'Don't know' and so 6 proceeded to answer the following question, rating database security practices against Agile software development. The results are given below:

Q10. Thinking about software development, please rate how well or otherwise you feel the following practices are applied in your organisation.

	Values						Percentages					
	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total	Not applied	Occasionally applied	Normally applied	Always applied	Don't know	Total
Native security by design (authentication within the	0	2	1	1	2	6	0.0	33.3	16.7	16.7	33.3	100.0
Use of domain accounts for application data access	2	2	2	0	0	6	33.3	33.3	33.3	0.0	0.0	100.0
SQL injection prevention techniques	1	0	1	2	2	6	16.7	0.0	16.7	33.3	33.3	100.0
Two-factor authentication	2	2	1	0	1	6	33.3	33.3	16.7	0.0	16.7	100.0
Principle of least privilege	1	1	2	1	1	6	16.7	16.7	33.3	16.7	16.7	100.0
Information silos (separation of e.g. customer, department data	1	2	0	1	2	6	16.7	33.3	0.0	16.7	33.3	100.0
Bug identification (for security issues)	1	1	2	0	2	6	16.7	16.7	33.3	0.0	33.3	100.0
					Answered	6					Answered	33.3
					Skipped	12					Skipped	66.7

Figure 4.6: Results from Question 10

Question 11 was a qualifying question on whether the respondents' organisations complied with the ISO 27001 standard. The 4 respondents who answered 'Yes' were directed to the next question on the perceived effectiveness of the ISO 27001 standard.

Concerning the question on new security technologies, the results are presented in rank order by popularity (number of 'Yes' answers) with the respondent metrics alongside.

Table 4.7: Results from Question 12

Rank	Technology	Responded 'Yes'	Responded 'No'	Responded 'Don't know'
1	Automated vulnerability reporting	8	0	0
2	Two-factor authentication	7	1	0
3	Single sign-on (abolition of application-level accounts)	6	1	1
4	Machine learning / AI systems for security control	6	2	0
5	Separation of security mechanisms from the database software	6	1	1
6	User behaviour modelling for threat/intrusion detection	6	0	2
7	Biometric user identification	3	4	1
8	Blockchain technology integration for security improvements	1	3	4
9	Negative' databases (obfuscation of data through false records)	1	5	2

The following question requested freeform answers on which other technologies the respondents would like to see implemented in database platforms. There were 3 responses:

> *"Native integration with secrets management tools"*
>
> *"Better integration with NoSQL"*
>
> *"Automated disconnection after a certain period of inactivity, without the possibility of saving passwords."*

Question 15 asked about whether better standardisation of data security controls across industries would be welcomed and invited freeform comments as responses. The 6 responses are given below:

> *"Yes, I believe standardization would be beneficial in that it would provide a known standard to follow and remove some of the "I think this is good enough" views toward security."*
>
> *"Yes as a base model to help enforce standards in businesses/industries where they do not exist – but could expose some weaknesses where not followed properly."*
>
> *"Standards should always be applicable. Most of them are not practicable."*
>
> *"Potentially, although this creates a single point of failure if not done perfectly."*

"Yes"

The following question asked if data security policies or procedures at the respondents' organisations had room for improvement. 8 responses were forthcoming of which 3 were 'Agree' and 5 were 'Strongly agree'.

The next question asked respondents for freeform answers suggesting ways in which data security could be improved at their organisations. 7 answers were received:

> *"More auditing/review, password rotation, elimination of shared app credentials, implementation of secrets management solution"*
>
> *"Not enforced, only approached when there is a flaw/hack…!"*
>
> *"We could push back more on vendors that require admin access for their application (although that risks vendor support in many cases). The policy of separate accounts for user/admin activities could be more robust."*
>
> *"There are a lot security problems [sic]. Database security is not the highest priority."*
>
> *"We have multiple legacy systems with poorly designed databases that we are unable to alter as software is 3rd party, including non hashed user password tables with root access."*

Question 18 asked how useful a central reference source on best practices in data security implementation would be to the respondents. This question was worded to tie into the first key deliverable of this research, the Data Security Framework, to establish if this is a viable and useful option. There were 7 responses to this question of which 3 answered 'Very useful', 3 answered 'Quite useful', and 1 answered 'Somewhat useful'.

The final question, asking for any further comments, gathered no substantial responses.

4.1.2 Primary Research - Interviews

Three interviews were conducted, the first two with industry practitioners and the last with an academic specialist, as detailed in Section 3.2.2. The interviews lasted approximately 30 minutes, were conducted via telephone and Skype and were based on the pre-planned questions detailed in Section 3.2.2. They were recorded using audio-capture software.

Before the interviews started, the ethical considerations detailed in Section 3.4 were made.

Using the Boyatzis (1998) framework, key phrases, sentiments and opinion have been extracted from the audio recording of the interviews. Each item of data is presented under the heading of the question last asked. Being semi-structured, there was some deviation from the pre-planned questions and so the headings vary. The interview question bank can be found in Appendix A. The findings are presented and codified with aggregation and thematic analysis detailed in Section 4.2.2. The data is an extraction of the participants' key statements, of which the critical elements are highlighted in yellow. The full table of interview outcome data and analysis is in Appendix B.

4.1.3 Secondary Research – Literature Review

As described in Section 3.2.3, the literature review has been ongoing since the beginning of the project and the outcomes have been collected and organised using an adaption of the grounded theory approach.

These were arrived at using the grounded theory adaption. In practice this was achieved using several methods – a spreadsheet program, to record and link data; pen-and-paper diagrams, to link concepts and draft themes; software tools for flowchart creation including draw.io; and the use of sticky notes and a whiteboard, in the manner of the Kanban task management system (Sugimori et al., 1977). Over 350 individual sources of information (academic and industrial) were considered, of which approximately 150 contributed directly to the data collected.

This secondary research, together with the outcomes of the primary research methods, therefore forms the fundamentals of the DSF. Chapter 2 contains a narrative literature review incorporating many of the findings, with Chapters 4 and 5, containing a description of the content and structure of the DSF, containing much of the detail. The high-level categories of the DSF were derived directly from the categories found through the literature review and are presented in Section 4.3.2, figure 4.16.

4.2 Data Analysis

4.2.1 Primary Research - Questionnaire

Analysis of the questionnaire results was carried out using standard statistical techniques within Microsoft Excel, which included percentage calculation, calculation of p-values to test statistical significance, and summation of various attributes.

The engagement of respondents with the questions is illustrated in Figure 4.1 and follows the same general shape as a sales funnel (Townsend, 1924). The term 'respondent attrition' refers to the number of respondents who abandoned the survey as it progressed from beginning to end, and this is visualised in the right-hand section of the diagram.

	Track Engagement		
Question #	*Industry*	*Academic*	*Total*
1 and 2*	17	1	18
3	16	0	16
4	14	0	14
5	11	0	11
6	11	0	11
7	9	0	9
8	8	0	8
9	8	0	8
10	6	0	6
11	8	0	8
12	4	0	4
13	8	0	8
14	3	0	3
15	6	0	6
16	8	0	8
17	7	-	7
18	7	-	7
19	7	-	7
20	1	0	1

** Note: Questions 1 and 2 are combined as they determine which track the respondent follows*

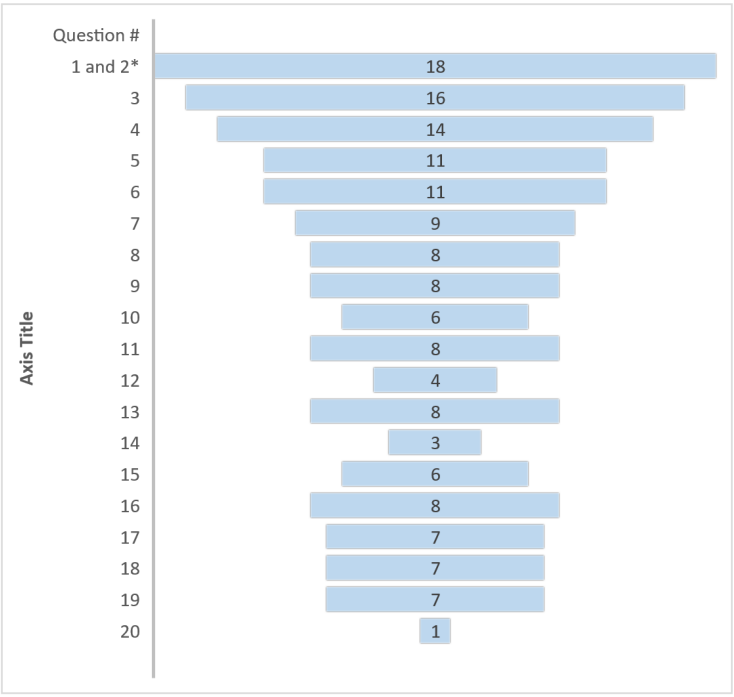


Figure 4.8: Questionnaire respondent attrition

In terms of the question contents, data collection yielded responses only along the industry track. This is symptomatic of the audience demographic at which the questionnaire was aimed, as academics constituted a small percentage of the potential respondent pool.

Qualifying Questions and Engagement

As detailed in Figure 4.1 the number of responses decreased, and the number of questions skipped increased as the questionnaire progressed. This attrition could have several underlying factors. It might indicate that the questionnaire was too long, or too complex, to hold interest, or it might be symptomatic of the nature of multi-tasking in that the respondents' attention spans were exhausted during the process, or the attrition might be the natural result of the lack of respondent's commitment to finishing the questionnaire due to having no stake in the

result. The low response rate weakens the validity of any conclusions made when the data is considered in aggregate and could have been rectified in the pilot stage if the pilot had been extended to more individuals or if more detailed feedback had been sought.

Viewpoints on Data Security

If the questionnaire was pitched at a non-specialist audience, one would expect the range of responses to follow a normal distribution, but the strong response in the 'Very important' category indicates that this subject is important to the respondents and is a strong indicator that the audience for the questionnaire was correct.

Applicability of Controls to Organisations

The first 5 non-qualifying questions in the industrial track dealt with how various practices are applied within the respondent's organisation. The practices are split into 5 categories across Questions 4-8: data access control, confidentiality, integrity, availability and overall data management. These categories resulted from the notes made during secondary research into the overall structure and aims of data management.

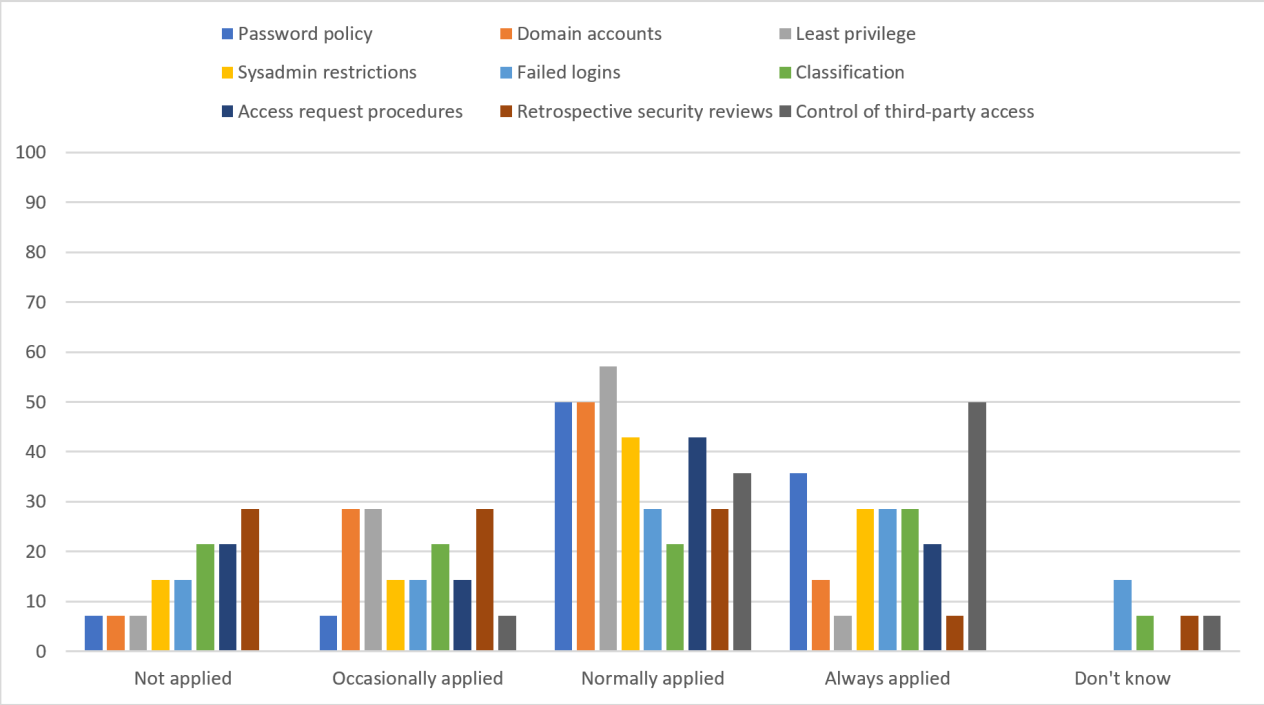


Figure 4.10: Visualisation of results from Question 4

Figure 4.10 illustrates the responses to Question 4, which seeks to discover whether various methods to improve security (as discovered during the literature review) are applied in practice. This question yielded some important statistical facts, which have been used to inform the interview questions. These include:

- 14.3% of respondents claim their organisation does not audit failed login attempts to the data platforms.
- 28.6% of respondents claim their organisation does not review user permissions regularly once applied.
- Ranking these data access controls, 92.8% of respondents claim their organisation has controls (applied occasionally, normally or always) on third-party access to their organisational data (inferring 7.2% do not). In contrast, 85.8% of respondents claim their organisation (occasionally, normally or always) restricts sysadmin access, which is an internal control (inferring 14.2% do not). This might indicate that internal threats are underrated compared to external threats.

There are some potential inconsistencies with these answers. The 14.3% of respondents who indicated no failed login auditing, for example, may be unaware that failed logins are automatically (by default) recorded in the error log and application log respectively of the two largest (by market share) RDBMS platforms on the market. However, the response may serve as a potential indicator of how important the respondent feels the issue is and may serve some purpose to confirm internal consistency of questions.

These findings, although questionable, constitute limited evidence of the disparity (referenced in the Research Aim) between the implementation of controls in organisations and the standards, best practices and academic research that make up the body of literature on how data security should be implemented.

These results would indicate that a significant minority of respondent's organisations are not implementing some of the core principles specified in the standard. This is backed up by the answers given to Question 12, 'How effective is the ISO 27001 standard in helping your organisation plan and manage their IT security?'. The respondent figure dropped to 4, but of the 4, 1 answered 'Not effective at all', 2 answered 'Somewhat effective', and 1 answered 'Very effective'. Plotting these answers against a normal distribution curve with 4 members across the 5 non-neutral categories and the respondent numbers on the Y-axis, as shown in Figure 4.11 we can see there is a skew towards a pessimistic viewpoint. However, the impact of a small sample size cannot be discounted when considering the cause of this skew.

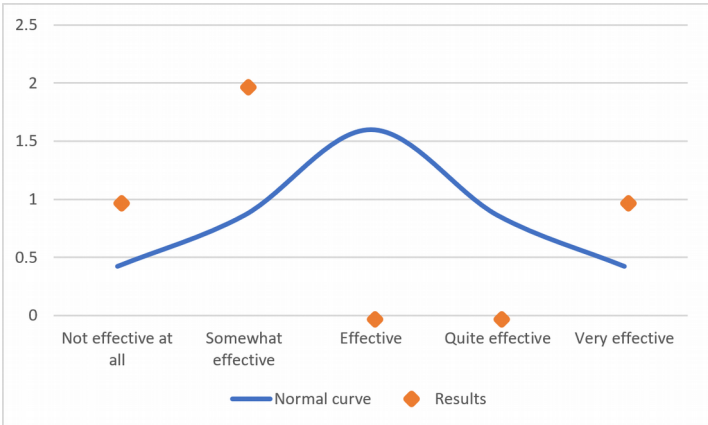


Figure 4.10: Stated effectiveness of ISO 27001 against the normal distribution curve

If the effect of small sample size is overlooked, we can test for statistical significance to ensure this perceived skew is quantitatively accurate. We can test this in Excel against these 4 results and the normal distribution curve by using a 2-tailed paired T-test, using the Excel function T.TEST(). Given the threshold of the null hypothesis (there is no significant difference) as $p = 0.05$, as is standard, we calculate $p = 0.94$ with these results. This indicates there is a statistically significant difference between a normal distribution and the results obtained. However, 4 is a small sample size and skew may naturally follow from the lack of responses. More evidence is required to prove the disparity viewpoint.

There was also demonstrated in several questions a significant deviation in security implementations from well-understood security practices. 27.3% of respondents claim that no regular checks are made for database corruption, while 18.2% claim no auditing takes place on business-critical data. The same percentage also claim there is no disaster recovery planning in place. While it is encouraging that these are minorities, we may conclude that there is a gap between best practices as understood from the literature review and the implementation of integrity-related controls in practice.

The results showed that on the whole controls for ensuring database availability are largely in place, and so does not concord with the disparity hypothesis. 100% of respondents claim server-level monitoring is at least occasionally applied with 88.9% claiming database-level monitoring is also at least occasionally applied. Despite 18.2% of respondents claiming lack of disaster recovery, 88.9% of respondents claim some application of automatic failover capabilities, showing that DR/BC has been considered. This raises the question of the trustworthiness of the respondents' answers as these two results are contradictory. One potential explanation is the disconnection made between DR/BC (the policy) and automatic failover as a control.

In terms of the effectiveness of overall data management, respondents claimed in 37.5% of cases that no ISMS was in place at their organisation. This result roughly correlates to the ISO 27001 compliance question, which requires an ISMS and had 2 negative answers of 8 (with 2 neutral). The regular review of security documentation is also required under ISO 27001 and had a correlating answer. There was a higher proportion of neutral answers to this question, possibly indicative of lack of knowledge on how data is managed more generally at the respondents' organisations. This could indicate a lack of awareness which may affect the trustworthiness of the result set in general.

Concerning Agile software development methodologies and data security practices, the results were mixed and showed no immediate positive or negative correlation.

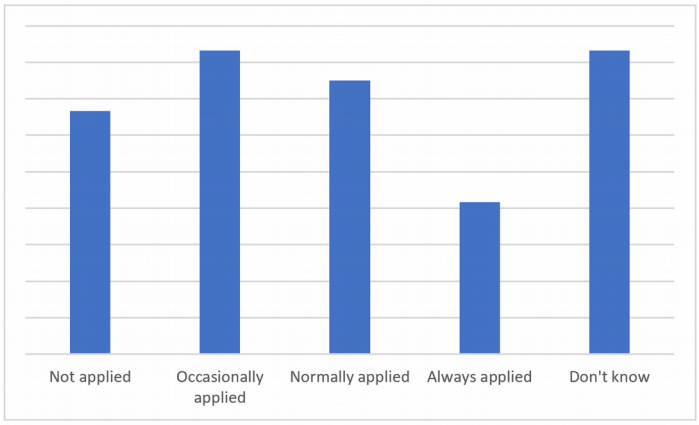


Figure 4.11: Aggregate response distribution for Question 10

Table 4.12: p-value calculation on the aggregate responses to Question 10

Category	Not applied	Occasionally applied	Normally applied	Always applied	Don't know
Results	133.4	166.6	150	83.4	166.6
Normal	70	140	280	140	70
p	0.059	0.457	0.002	0.115	0.016
Significant	Yes	No	No	Yes	No

These values show that by majority verdict across the categories, there was no statistical significance and so no firm conclusions can be safely drawn.

New Technologies

Question 12 presented a list of potential new technologies that could have applications in data security. This list was sourced from the reading and review carried out during secondary research, and so this question acts to triangulate upon how viable other respondents to the questionnaire believe each technology could be.

The first result may indicate that respondents feel there is a disconnection between threat identification and vulnerability detection within data platforms and is used as a theme for further detail in the subsequent interviews. The second and third-ranking options both concern user access control. Coupled with the freeform answers in the following question, this shows a potential gap in an ideal standard for user access control and actual implementation – ideas from the respondents included integration with password management tooling.

Standardisation and Improvements

The freeform answers to Question 15 all replied in the affirmative to the question on whether standardisation of data security controls would be beneficial, with some qualifiers. One comment which is of particular interest to bring to interview is 'Standards should always be applicable. Most of them are not practicable'. This could be indicative of a wider lack of awareness of how standards should be applied, and perhaps identify gaps (mapping to O5, the delivery of a GAP analysis). This corresponds to the findings from the interview stage.

Reference Model

This question aimed to determine how useful the respondents feel a reference model (O3, key deliverable) would be on a 4-point Likert scale with a neutral ‘Don’t know’ answer. 7 respondents answered affirmatively (0 responses for ‘Not useful at all’ with 0 ‘Don’t know’). This provides some assurance that the Data Security Framework could provide value for these practitioners.

The insights from analysing the results of the questionnaire have been summarised from the discussion above and added to Figure 4.14 below, mapping questions to objectives and displaying the results.

Question	Pathway	Summary	Category	Purpose	Measurement methodology	Scale	Maps to objective	Insights from Analysis
INTRO	Both	Introduction / Welcome screen	n/a	Landing page	n/a	n/a	n/a	-
1	Both	Practitioner?	Validate participant	To determine if the participant is an active practitioner or a researcher/teacher.	Quantitative	Binary	O2	• The number of respondents to the survey is smaller than anticipated. • There is evidence to suggest the survey may have been too complex or time-consuming to sustain interest by respondents in completing all questions. • A significant minority of respondents claim their organisation do not apply controls to audit failed login attempts, review user permissions or monitor 3rd party data access, which implies serious non-adherence to industry standards and best practices. • More respondents claim their organisation restricts external data access than internal data access, indicating that external and internal threats are not treated with the same degree of controls. • The evidence suggests that the disparity hypothesis is true, proved by the existence of the minority and the p-value calculated in relation to Question 4. The exception to this is with regards to controls for availability, to which respondents answered positively in all cases. • More than a third of respondents claim that ISO 27001 adherence is not in place at their organisation, implying that either their organisations are exempt (or believe themselves to be so); or do not see the value in becoming compliant with the standard. This result may also indicate that the respondents are not aware of their organisation's compliance status. • The results from questioning around data security in software engineering were inconclusive.
2	Both	Academic?	Validate participant	To determine if the participant is an active practitioner or a researcher/teacher.	Quantitative	Binary	O2	
3	Practitioner	Importance of data security to self	Assess data security importance	To establish the relative importance of data security to the participant, for group aggregation	Quantitative	Likert	O2	
4	Practitioner	Importance of data security to organisation	Assess data security importance	To establish the relative importance of data security to the participant's organisation, for group aggregation	Quantitative	Likert	O2	
5	Practitioner	Rate knowledge of data access controls	Assess applied data security controls	To determine how familiar the participant is with a range of different data access controls, to establish specific ones which aren't in common use for triangulation via questioning in interview	Quantitative	Likert	O2, O4	
6	Practitioner	Rate knowledge of controls for data confidentiality	Assess applied data security controls	To determine how familiar the participant is with a range of different controls aimed at maintaining confidentiality, to establish specific ones which aren't in common use for triangulation via questioning in interview	Quantitative	Likert	O2, O4	
7	Practitioner	Rate knowledge of controls for data integrity	Assess applied data security controls	To determine how familiar the participant is with a range of different controls aimed at maintaining data integrity, to establish specific ones which aren't in common use for triangulation via questioning in interview	Quantitative	Likert	O2, O4	
8	Practitioner	Rate knowledge of controls for availability	Assess applied data security controls	To determine how familiar the participant is with a range of different controls aimed at maintaining data management, to establish specific ones which aren't in common use for triangulation via questioning in interview	Quantitative	Likert	O2, O4	
9	Practitioner	Rate knowledge of controls for data management	Assess applied data security controls	To determine how familiar the participant is with a range of different controls aimed at maintaining availability, to establish specific ones which aren't in common use for triangulation via questioning in interview	Quantitative	Likert	O2, O4	
10	Practitioner	Agile qualification	Assess applied data security controls	Qualification question to determine if the participant's organisation uses the Agile software development methodology	Quantitative	Ternary	n/a	

Figure 4.13(a): Questions mapped to objectives mapped to insights from result analysis

11	Practitioner	Rate knowledge of controls for data security in the context of software development	Assess applied data security controls	To determine how familiar the participant is with a range of different controls aimed at data security in the context of software development practices, to establish specific ones which aren't in common use for triangulation via questioning in interview	Quantitative	Likert	O2, O4	
12	Practitioner	ISO 27001 qualification	Assess opinion on ISO 27001	To establish whether the participant's organisation is subject to ISO 27001 certification.	Quantitative	Ternary	n/a	
13	Practitioner	Rate the perceived effectiveness of ISO 27001 on the participant's organisation's ISMS	Assess opinion on ISO 27001	To establish whether the participant feels ISO 27001 has had a positive net benefit - may supply justification for an auxilliary framework in the form of the DSF	Quantitative	Likert	O2, O5	
14	Practitioner	Which security technologies would the participant like to see implemented in databases?	Assess views on potential future database technologies	To rank the popularity of a range of alternative security technologies with a view to present the most popular as research directions in the DSF.	Quantitative	Ternary	O2, O4	• There may be a disconnection between threat identification and vulnerability detection in data security implementations .
15	Practitioner	Participant to suggest new security technologies.	Assess views on potential future database technologies	To gather new ideas on potential directions for database security techniques for use in the DSF.	Qualitative	Open	O2	• There is evidence of differences between best practice and implementation of user access control.
16	Practitioner	Whether standardisation of security controls across platforms is worthwhile	Assess views on policy and procedures	To determine whether the participant favours a heterogenous or homogenous approach to security, which may inform analysis on why security standards are divergent.	Qualitative	Open	O5	• Data professionals may have no influence over systems maintained by 3rd parties.
17	Practitioner	Qualification, whether the participant feels the ISMS policies and procedures at their organisation can be improved	Assess views on policy and procedures	To establish whether the participant feels there is a gap between what should and is implemented.	Quantitative	Likert	O5	• It is felt that database security is not the highest priority in the context of a wider security landscape.
18	Practitioner	What suggestions the participant has for improving data security policies and procedures in the ISMS.	Assess views on policy and procedures	If the answer to Q17 is on the positive side of the neutral answer, find out more information on exactly what ideas the participant might have. Could be triangulated in subsequent interviews.	Qualitative	Open	O2, O4	• More auditing of logins and password management could be required. • There is room for the integration of password management applications into user access control.
19	Practitioner	Proposition defining the DSF and asking whether it might be useful in practice.	Assess views on policy and procedures	Gain a view on whether the DSF is likely to be used in practice if fully developed - adds to justification and impact on wider community.	Quantitative	Likert	O5	• All respondents felt a database security reference model would be beneficial, which adds weight to the viability and usefulness of the DSF (objective O3).
20	Both	Any other comments to make about data security?	n/a	Gather any thoughts or comments that the participant has that might have some bearing on the subject	Qualitative	Open	O2, O4, O5	-
HARD EXIT	Neither	This will be invoked if the participant answers 'no' to both questions 1 and 2, since they will not be the target demographic for this survey and may not be able to provide answers	n/a	Exits the survey politely without asking further questions.	n/a	n/a	n/a	-
SOFT EXIT	Both	This is invoked at the end of the survey once all questions have been completed, with a thank you message to the participant.	n/a	Exits the survey.	n/a	n/a	n/a	-

Figure 4.13(b): Questions mapped to objectives and insights from result analysis (continued)

4.2.2 Primary Research - Interviews

Thematic analysis (Boyatzis, 1998) is used to analyse the data following the staged methodology set out by Braun, Clarke and Terry (2012) – in essence, this is to divide the *corpus* of interview data into key statements, codify these statements and group them into themes. From the codes, themes and interview data, construct statements, or conjectures, which correspond to the data collected. This is also a useful method of identifying classes of problem to investigate (the themes). This was implemented as follows: for input (the data) the statements extracted from the interview in Section 4.1.2 are dissected into 'codes', or specific concepts. The first stage of this, assembling these concepts, is shown in the data collected and presented in Section 4.1.2.

Next, these concepts were assembled into themes. These themes are informal and encompass the concepts. The findings have been divided into 6 major themes. Each theme, or component of a theme (if more appropriate) contains similar or co-dependent ideas:

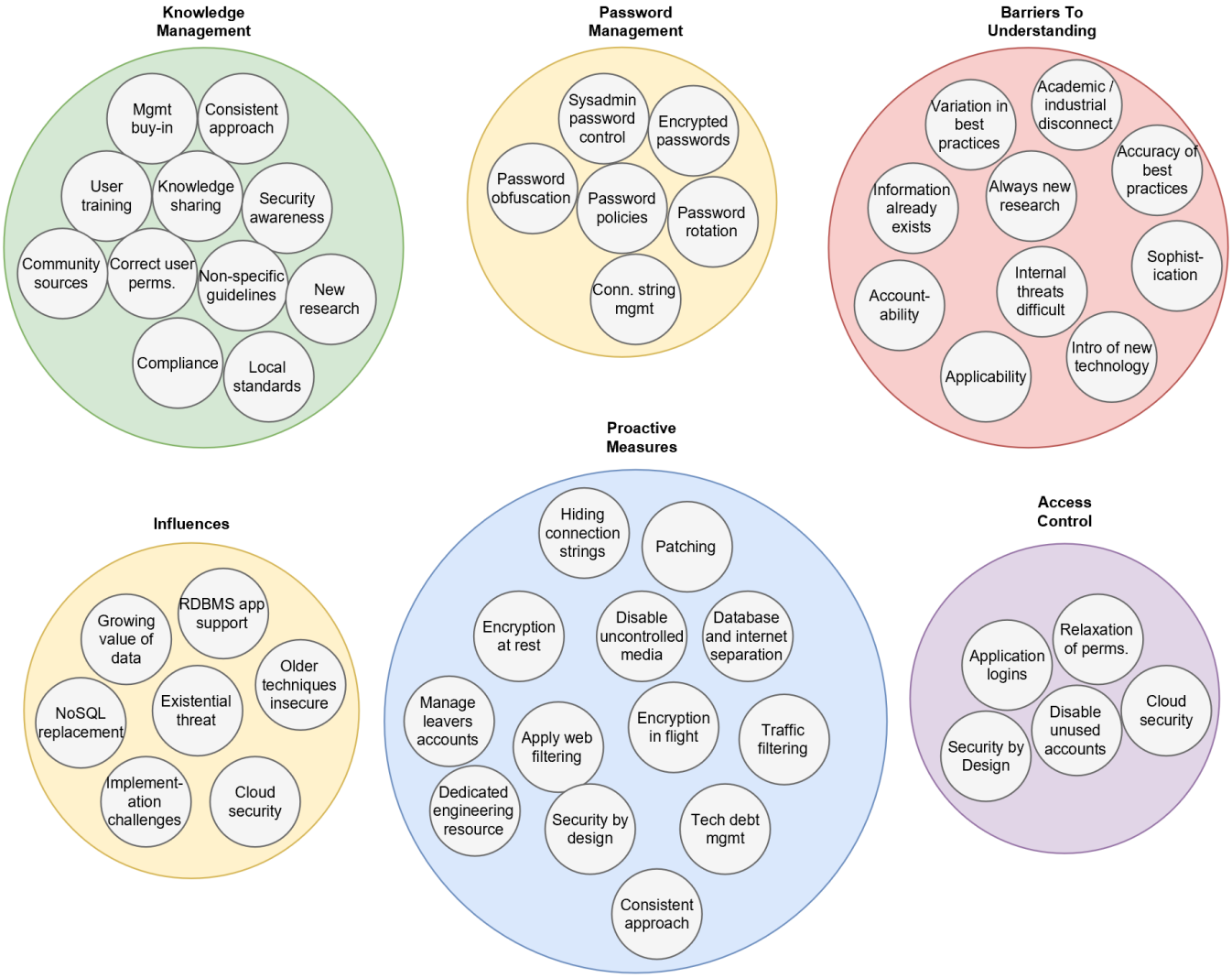


Figure 4.14: Thematic grouping of the interview outcomes

Finally, the themes are expressed as conjectures, in the same manner as the insights expressed in the questionnaire analysis. Each conjecture is a statement of opinion which aggregates and restates the data collected in the interviews, sometimes across themes. This technique became more powerful as more interview data was added. The themes and conjectures are shown in Figure 4.16. These conjectures contributed to the content of the associated topics in the DSF and helped provide direction for the GAP analysis.

The following conjectures are a sample of the full list, which can be found in Appendix C.

Table 4.15: Emergent conjectures from interview analysis

Theme / Concept	Insight
Knowledge management / consistent approach	A consistent approach can strengthen the effectiveness of good password management.
Knowledge management / security awareness	Better security awareness can aid adherence to password policies
Knowledge management / non-specific guidelines	Having vague guidelines can cause implementation challenges.
Knowledge management / non-specific guidelines	Having vague guidelines causes variation in best practices within an organisation, creating vulnerabilities.
Knowledge management / correct user permissions	Setting the correct user permissions is associated with good user access control.
Barriers to understanding / accountability	To ensure accountability, put in place effective access controls. Likewise, a drive for accountability promotes good access control mechanisms.
Barriers to understanding	Barriers to understanding lower the effectiveness of proactive measures to safeguard data security.

4.2.3 Secondary Research – Literature Review

The secondary research was conducted using techniques derived from grounded theory, where extensive memos and references on each of the topics described in Section 4.1.3 became outputs and culminated in the production of the DSF and GAP analysis, combined with the insights from the primary research. The narrative version of this literature review fulfils objective 1 (O1) and is presented in Chapter 2 of this document. The insights from the literature review have also been wrapped into the DSF as subject entries against each concept.

The codification of the literature review led to a large list of data security topics which is presented in Section 4.3 below, and Figure 4.17. Each item in this list forms a category in the Data Security Framework. As described elsewhere the structure of the DSF is such that this list is effectively the root level of the taxonomy, or hierarchy, of data security topics that were researched.

4.3 Interpretation in relation to the objectives

4.3.1 Literature Review and Primary Research

O1. Conduct a literature review synthesising and summarising the progress made in database security research;

This objective has been achieved through the production of the review in Chapter 2, together with the content of the DSF which is detailed under O3, below.

O2. Conduct a comprehensive search of the practitioner literature and conduct primary research by way of semi-structured interviews and a survey to determine a collection of best practices in database security;

In total, one questionnaire and three interviews were carried out together with the literature review of Chapter 2 and Objective O1. Using statistical analysis of the questionnaire results to infer conclusions, and using thematic analysis of the interview data supplied in Appendix B, insights were derived from the results (these are given for the questionnaire in the final column of Figure 4.13(a) and (b), above, and for the interviews from the final column of Figure 4.15, above). This, together with the codified notes against each topic, were used to create the DSF. For example, item 2 in Figure 4.15 ('better security awareness can aid adherence to password policies') directly led to the recommendation in items 2.15 and 2.16 of the DSF that the password policy of the organisation should be circulated to all staff members.

4.3.2 DSF Structure

O3. Consolidate the findings of 1 and 2 to produce a Data Security Framework which will be a detailed taxonomy, or catalogue, of database security concepts, best practices, methods and techniques together with their linkages, contextual information and examples;

The following section deals with the construction of the DSF and is broken into two sub-sections, structure and content. Implementation examples are discussed and provided in Chapter 5. A fuller example of the DSF structure including a topic record is available in Appendix D.

When considering the DSF the following definitions are given:

- Category: A grouping of one or more concepts into a single common theme
  - For example, 'Development' incorporates three concepts
- Concept: A single topic, idea or theory in the domain of database security belonging to a category
  - For example, 'Enforcing security-by-design' belongs to 'Development'

A category has the following attributes:

- ID: A unique numerical one-part identifier in the form *x*, where *x* is a positive integer greater than 0.

- Title: The title of the category
- Precis: A short summary of the category, typically a few sentences.

A concept has the following attributes.

- ID: A unique numerical two-part identifier in the form *x.y*, where *x* is the parent category identifier and *y* is a positive integer greater than 0.
    - For example, 2.9 or 3.13.
- Title: The title of the concept
- Precis: A short summary of the concept, typically a few sentences, relevant to database security.
- Threats: A description of the attack surface that relates to the concept (if any); alternatively, how the threat profile is affected.
- Best practices: A brief description of best practices that relate to the concept.
- Academic summary: A brief description of any relevant academic research in this area.

Therefore, the structure of the DSF can be visualised as shown in Figure 4.16:

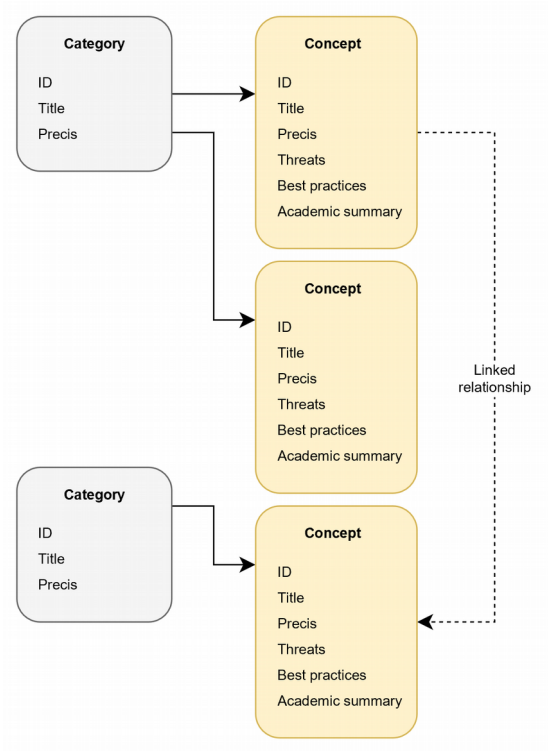


Figure 4.16: Visualisation of the DSF (partial)

4.3.3. DSF Content

Using the outcomes of the literature review and the outcomes from O1 and O2, the high-level list shown in Figure 4.17 shows the unique identifiers, categories and concepts of the DSF.

There are 16 categories in total, with 114 concepts, totalling approximately 730 data points. ID is typed as described, in the format X.Y; Title, Precis, Threats, Best Practices and Academic Summary are text data. Within the text data, terms which relate to other concepts can be linked. The exact method of linking depends on the implementation – in the relational implementation, a separate table is required linking concepts together. In the non-relational implementation, a nested set of values provides linked concepts (mono-directional) from each record. In the .HTM implementation, all links are hyperlinks using HTML or marked as Related Topics. Implementations are discussed further in Chapter 5 and three example implementation frameworks are provided.

UQID	Category	Concept
1.1	Abstract concepts	Bell-LaPadula model
1.2	Abstract concepts	Biba model
1.3	Abstract concepts	Clark-Wilson model
1.4	Abstract concepts	GRC Framework (Hill model)
1.5	Abstract concepts	State machine security representation
2.1	Access control	3rd-party access regulations
2.2	Access control	Authorisation administration
2.3	Access control	Avoiding use of basic authentication
2.4	Access control	Centralised administration
2.5	Access control	Content-based access control
2.6	Access control	Context-aware control
2.7	Access control	Discretionary access control
2.8	Access control	Incorporation of user identity characteristics
2.9	Access control	Intrusion detection mechanisms
2.10	Access control	MAC - Bell/LaPadula - No read-up, write-down
2.11	Access control	Mandatory access control
2.12	Access control	Name-based access control
2.13	Access control	Ownership administration
2.14	Access control	Password management
2.15	Access control	Password policy
2.16	Access control	Polyinstantiation
2.17	Access control	Principle of least privilege
2.18	Access control	Proactive intruder detection
2.19	Access control	Restriction of sysadmin accounts
2.20	Access control	Retrospective review
2.21	Access control	Separation of duties
2.22	Access control	Temporal factors

UQID	Category	Concept
3.1	Auditing	Audit changes to data
3.2	Auditing	Audit changes to schemas
3.3	Auditing	Audit failed logins
4.1	Availability	Cloud databases (DaaS)
4.2	Availability	Database unavailability
4.3	Availability	DoS / DDoS protection
5.1	Business view	Agile development methodology
5.2	Business view	Disaster recovery planning
5.3	Business view	Formal authorisation procedures for new/revised access
5.4	Business view	Insurance
5.5	Business view	Legacy system security requirements
5.6	Business view	Move to proactive rather than reactive security
5.7	Business view	News dissemination through external sources
5.8	Business view	Operational recovery planning
5.9	Business view	Regularly review ISMS documentation
5.10	Business view	Risk management
5.11	Business view	Security incident management response procedures
5.12	Business view	Standards not practicable
5.13	Business view	User training
6.1	Confidentiality	Air-gapped systems
6.2	Confidentiality	API/key-based access control
6.3	Confidentiality	Data mining
6.4	Confidentiality	Hippocratic' (Agarwal) databases
6.5	Confidentiality	Multi-level relations
6.6	Confidentiality	Negative authorisation (revoke)
6.7	Confidentiality	Password rotation
6.8	Confidentiality	Row-level filtering

UQID	Category	Concept
6.9	Confidentiality	View-based access
6.10	Confidentiality	Virtual private database
7.1	Data categorisation	Data labelling
8.1	Development	Developer training in data security
8.2	Development	Enforcing security-by-design
8.3	Development	Separation of production and test data
9.1	Encryption	Certificate/key management
9.2	Encryption	Common Criteria Certification
9.3	Encryption	Different types of cryptography
9.4	Encryption	Encryption-at-rest
9.5	Encryption	Encryption-in-flight
9.6	Encryption	FIPS 140-2 standard
9.7	Encryption	Use of standards
10.1	Environmental	3rd-party security requirements (software/DBs)
10.2	Environmental	BYOD policies
10.3	Environmental	Data-sharing policies for staff
10.4	Environmental	Misinformation
10.5	Environmental	Patching and updates
10.6	Environmental	Physical security
10.7	Environmental	Pre-existing documentation
11.1	Exploitation	CSV injection
11.2	Exploitation	Homoglyphic attacks
11.3	Exploitation	Output leakage
11.4	Exploitation	SQL injection
12.1	Integrity	Analysis of behaviour
12.2	Integrity	Anti-corruption measures
12.3	Integrity	Data completeness

UQID	Category	Concept
12.4	Integrity	Data quality
12.5	Integrity	Digital signatures
12.6	Integrity	Logging
12.7	Integrity	Loss of trust in data
12.8	Integrity	Security of backups
12.9	Integrity	Semantic integrity
13.1	New techniques	Automated disconnection forced from RDBMS
13.2	New techniques	Automated vulnerability reporting
13.3	New techniques	Biometric authentication
13.4	New techniques	Blockchain integration
13.5	New techniques	Captcha-style authentication
13.6	New techniques	Cross-platform security integration e.g. NoSQL
13.7	New techniques	Integration of SSO
13.8	New techniques	Integration with password management tools
13.9	New techniques	Machine learning-led / AI-led security analysis
13.10	New techniques	Separation of security from DB mechanisms
13.11	New techniques	Two-factor authentication

UQID	Category	Concept
14.1	Obfuscation	Connection string obfuscation
14.2	Obfuscation	Data masking
14.3	Obfuscation	Negative databases
14.4	Obfuscation	Privacy concerns
15.1	Role-based authentication	Application roles vs. user roles
15.2	Role-based authentication	Automated role management
15.3	Role-based authentication	Object-based
15.4	Role-based authentication	Task-based
16.1	Standardisation/compliance	COBIT
16.2	Standardisation/compliance	Compliance to internal policies
16.3	Standardisation/compliance	GDPR
16.4	Standardisation/compliance	Insufficient control specifications
16.5	Standardisation/compliance	ISO 27001
16.6	Standardisation/compliance	NIST 800-53
16.7	Standardisation/compliance	PCI-DSS
16.8	Standardisation/compliance	Sarbanes-Oxley

Figure 4.17: List of categories and concepts summarising outcomes of the literature review, forming the structure of the DSF

4.3.4 GAP Analysis

O4. Produce a GAP analysis between the academic and industrial states of database security and propose a catalogue of findings for future research or implementation;

During the research, differences between the methods used in industry and the methods investigated in the literature became apparent. In general, older principles of information security, which are well-understood, are in common use. Examples of this include the use of password management policies and mandatory access control, both of which have a substantial presence in the literature and both of which are embedded into the feature sets of all major RDBMSs.

However, some topics have more presence in the practitioner domain and less so in the academic; and likewise, some topics (particularly newer topics) have more presence in the academic domain and are little-used in industry. This GAP analysis seeks to define some of the categories and topics for which the practitioner-academic split is unbalanced and comment upon potential applicability or research potential. This detailed commentary is augmented with a succinct SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis per domain where the applicability or research potential of each item is aggregated and summarised to produce a shortlist of recommendations for future industry applications and future research directions.

In the term GAP analysis, GAP stands for Good, Average, Poor, and is a term borrowed from the PRINCE2 project management methodology (Bentley, 2012).

We first place each topic into a Good, Average or Poor classification for each domain. The decision on where each topic is placed is based on the information collected during the primary and secondary research for the topic. For example, 'New Techniques / Biometric Authentication' is placed into the 'Poor' classification for industry (in the context of data systems) as no RDBMSs currently implement such security mechanisms, being reserved for some applications, and neither are there any developments in data typing that specifically supports biometric data. However, there are interesting examples of potential directions in research for improving biometric authentication systems for data access (Ratha, Connell and Bolle, 2001) and other related research available. In this instance, the topic would be rated 'Good' for the academic domain. This approach is repeated for each topic.

The relevant questions that produce the ranking are, for each domain:

Industrial:

In the industrial sphere, how well is the topic understood and implemented, as it applies to data security?

Academic:

In the academic sphere, how extensively is the topic covered in the literature?

The full list of ratings against each topic are supplied in Appendix F.

From this list, the table below shows the topics ranked as Poor - there is either poor understanding and application in the industrial domain, or a lack of coverage in the literature and potentially room for more research. It is recognised that these are subjective ratings, however these stem from the findings of the literature review, the questionnaire and the interviews which are based on the analysis presented in Chapter 3. Internal validity is provided by a correlation between some of the narrative comments received in the survey instruments and the observations from the literature, for example comments on the integration of password management applications to RDBMSs were reflected in the lack of research into this area and the absence of this feature from all major RDBMS platforms.

Please see Appendix G which contains the GAP analysis in full – 48 topics together with their description and applicability or research opportunities.

In Section 5.1.4, these GAP analysis outcomes are amalgamated into a set of conclusions using SWOT analysis.

4.3.5 Validation and Reflection

O5. *Reflect on wider factors such as cultural or technological trends, research and implementation challenges to help refine and validate the model, proposing future work.*

Through collecting and analysing the data from the survey instruments and the literature review, certain themes were evident. It appears that these themes have a strong influence on the opinions expressed especially in the outcomes of the survey instruments. Additionally, the data is influenced by the direction and popularity of certain technological developments and current prevalent directions in academic research.

These trends include:

- Machine learning and artificial intelligence in the most recent literature, with many examples of new features particularly in cloud platforms using this technology. This has a direct bearing on security since often these systems are integrated with data storage systems – e.g. Microsoft offers Azure Machine Learning which can integrate with Azure DB (relational) and Cosmos DB (non-relational) data stores. With ML / AI capturing the *zeitgeist* of current technological trends, the term ‘big data’ has come to be coined as a description of large-scale data processing requirements to support complex application needs. These come with security challenges including ensuring data integrity, enforcing encryption-at-rest on rapidly-updating data and restricting access to data in flight.

- There appears to be a degree of disillusionment evident in the opinions expressed by the interviewees in particular – in both the content and tone of the replies. The impression is left that there is a disconnect between practitioners and management, with practitioners having little understanding of the details involved in data governance and the managers having little appreciation for the complexity of implementing and administering security controls. Although this wasn’t backed up in the literature, this

is a dangerous mindset as it creates barriers in designing and delivering effective information security management systems.

- Cloud computing is a popular and recent phenomenon where systems and data ordinarily resident on-premises, that is on a physical server owned by the organisation, is now resident on virtual servers leased from a larger service provider and accessible only via network connection. This creates data security challenges such as ensuring encryption when data is being transferred over the open Internet (i.e. by creating VPNs or using public-key encryption via key pair). It also introduces the ability for malicious actors to intercept and replace communications e.g. via man-in-the-middle (MITM) attacks. While cloud computing is convenient a whole new set of challenges needs consideration, as evidenced by the existence of the dedicated ISO security standard for cloud computing, ISO 27017.

In terms of the Database Security Framework, the speed and nature of technological progress means such a model, if implemented, would require an update process to ensure it remained relevant, and maintenance by some central group (as Internet standards are maintained by organisations such as W3). Also, given that practitioners appear to prefer clear guidance on enforcing security controls, the DSF may be more useful expanded into platform-dependent variations, although at the risk of increasing complexity and crossing the divide between framework and technical manual.

Further validation of the model is also recommended since the content was assembled from the observations and statements that were synthesised and aggregated from wider literature and a large *corpus* of data. It is likely that a subjective viewpoint has produced outcomes which could suffer from researcher bias. To solve this, further validation via external experts would be highly beneficial.

However, despite shortcomings it is apparent that data security is a topic is increasingly relevant and topical due to the speed of technological change, the growing value of data and the growing cultural awareness of the importance of data control, particularly considering recent and repeated high-profile data breach incidents. The Data Security Framework and accompanying GAP analysis are, hopefully, contributions to the area with the view that future work can expand and correct such a framework, and they can be disseminated to a wider community to assist all with the aim of improving the data security landscape for all.

4.4 Interpretation in relation to the research aim

The results of the research produced an extremely large body of data which needed to be assessed, sorted, linked and reduced into a finite set of conclusions that fit into the Data Security Framework model and which were able to be evaluated to identify gaps between academic research and industrial applicability. This resulted in a broad range of categorisations which, it is felt, is representative of the variety of issues within the field of data security at large. It is not claimed that this represents an exhaustive summary of the material. However, the data produced was helpful

in producing the content for the DSF and GAP analysis, which is given in more detail by the analysis in Section 4.3.3 and Section 5.1.4. Interpreting this data to form a coherent product was a central aim of the research, and this aim has been met, although there is scope for improvement by adding further detail and validation.

4.5 Chapter Summary

In this chapter we presented and analysed the data collected using the methods specified in Chapter 3. We collated, aggregated and drew preliminary conclusions from the data, and demonstrated a range of analytical techniques. We constructed the Data Security Framework and presented the taxonomy of content. We produced a full list of outcomes for the GAP analysis. We also considered the wider technological and cultural landscape.

5. Conclusions

5.1 Conclusions about the objectives

5.1.1 Objective 1: Literature Review

The purpose of this objective was to conduct an extensive literature review into data security, examining both practitioner technical material and academic literature, to build a database (in the non-technical sense) of information that could be used to populate a framework for data security. This objective was met, although restricted somewhat by available time and complexity of the material. The outcome of the literature review was a list of topics and for each topic, extensive commentary, memos and linkages to other topics. This helped build a coherent picture of the entire subject, which the following analysis constructed into a taxonomy of data security topics and a GAP analysis of practitioner and academic domain differences. This taxonomy structure was presented in Sections 4.3.1 and 4.3.2. Given that the literature review process successfully produced the wide variety and depth of information required, this objective was successfully met.

5.1.2 Objective 2: Discover Best Practices

The objective was to uncover best practices through both literature review, primarily of practitioner literature, and through primary research in the form of survey instruments – questionnaire and interviews. The outcomes, especially of the interviews, were very helpful in validating some of the information in the literature review around best practices. This internal validity and additional detail were crucial in constructing the DSF and performing the GAP analysis, and much of the DSF content derives from insights from this stage. However, the amount of information to summarise and parse into actionable statements was substantial.

A limitation of the research methods chosen was a lack of balance between industry and practitioner sources. Although an interview was conducted with a suitable academic, this constitutes a single source of information with no triangulation between this and the questionnaire outputs, since the participants in the questionnaire were exclusively practitioners. This meant the only internal validation for the academic conclusions is against the published literature and is consequently weaker than the conclusions for the industrial domain.

5.1.3 Objective 3: Data Security Framework

This objective was about the creation of a Data Security Framework – the latter word, framework, to refer to the design rather than any potential implementations. However, the feasibility of implementing any designed system is important, and demonstrates the viability of the research outcomes and the practical applicability to practitioners

and academics, a key stated component of the objective. Although the implementation of the DSF is entirely non-platform dependent and exists as an abstract design, three possible implementations are presented here.

The first implementation is relational, where the structure has been normalised to third normal form and is presented as a UML entity-relationship diagram; the second implementation non-relational, where the structure is described in terms of a document and an example JSON syntax is provided; and the third as an actual implementation in a .HTM file, compatible with Windows Help in Windows 10, from which several screenshots are provided.

5.1.3.1 Relational Schema

The format of the DSF, described in Sections 4.3.1 and 4.3.2, has a schema. A schema has a distinct meaning in relational data theory, and is described succinctly by the Oxford English Dictionary (Oxford University Press, 2018) as:

"... a conception of what is common to all members of a class; a general or essential type or form."

This means that the schema encompasses all the properties of a topic record and we can store these in a relational format. In relational modelling, the properties are attributes, and the owning objects are entities. We use normalisation, a standard relational modelling technique that has some straightforward rules (Connolly and Begg, 2004), to create the following schema for the DSF which is compliant to 3rd normal form (3NF). Figure 5.1 below shows a simplified entity-relationship diagram, using UML, for the schema:

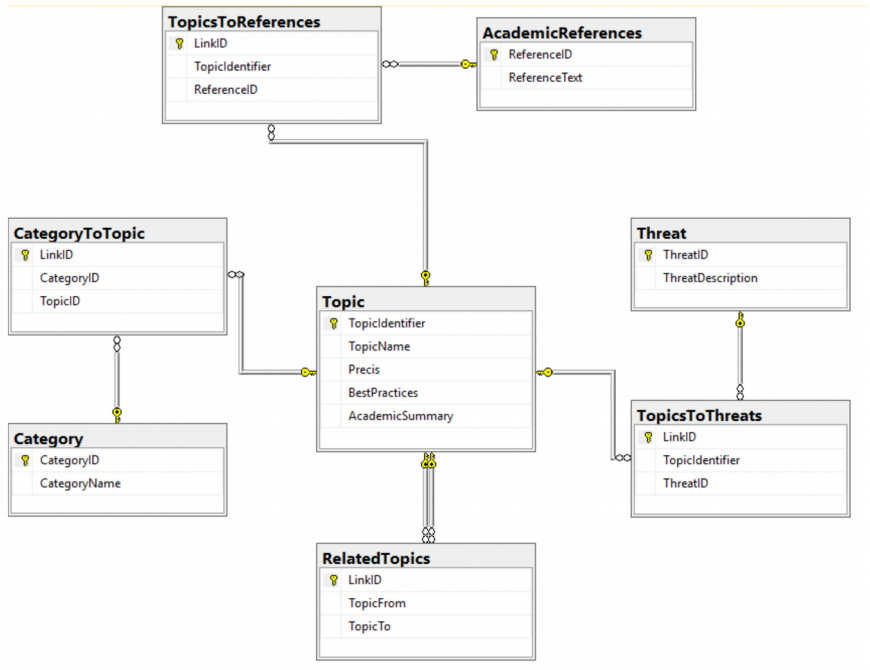


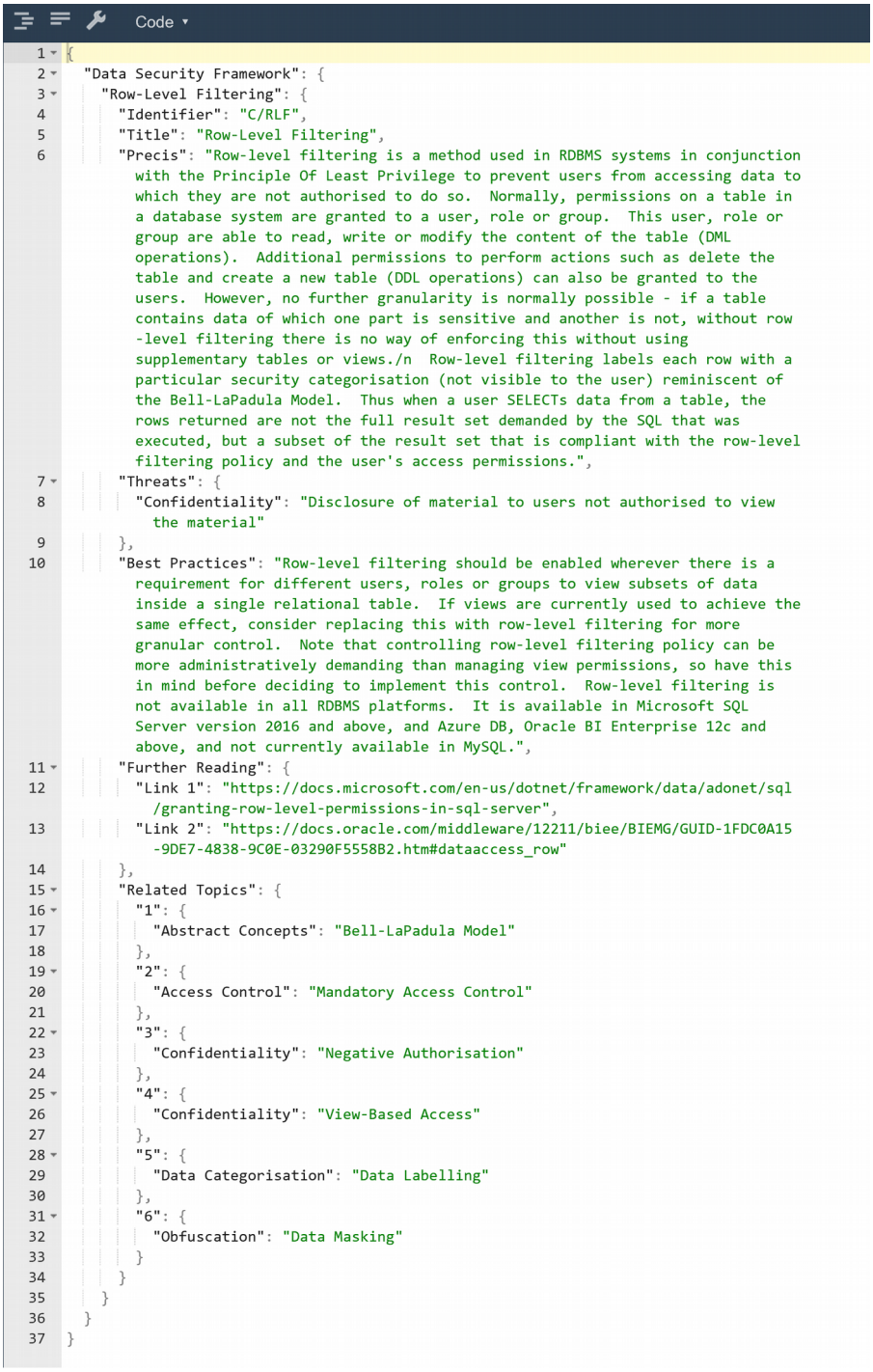
Figure 5.1: UML entity-relationship diagram for the DSF schema

The schema can be expressed in any RDBMS as a series of SQL DDL statements. The code in Appendix E shows one such implementation in Transact-SQL, the Microsoft SQL Server ANSI-SQL variant.

The schema can be populated easily through INSERTs into the relevant tables. In practice, one could create stored procedures to select from, insert to, update and delete from the tables.

5.1.3.2 Document Store (JSON)

Document stores are typically unstructured 'NoSQL' repositories of data that do not have set schemas – that is to say, each document within the store can have a different structure. This type of repository is useful for unstructured analytics data, for example, such as web traffic or for storing multimedia files. It is not typically used where the content is in a known schema, such as the DSF design, but as each topic in the DSF contains a substantial amount of narrative text, there is an argument for using unstructured storage for better indexing capabilities – this would enable faster searching, for example, than in relational stores. Given the potential practicality of a document-store based implementation, the following figure shows a single topic record in JSON, a popular format compatible with document stores including AWS Dynamo DB, Azure Cosmos DB and CouchDB.



```
{
  "Data Security Framework": {
    "Row-Level Filtering": {
      "Identifier": "C/RLF",
      "Title": "Row-Level Filtering",
      "Precis": "Row-level filtering is a method used in RDBMS systems in conjunction
        with the Principle Of Least Privilege to prevent users from accessing data to
        which they are not authorised to do so.  Normally, permissions on a table in
        a database system are granted to a user, role or group.  This user, role or
        group are able to read, write or modify the content of the table (DML
        operations).  Additional permissions to perform actions such as delete the
        table and create a new table (DDL operations) can also be granted to the
        users.  However, no further granularity is normally possible - if a table
        contains data of which one part is sensitive and another is not, without row
        -level filtering there is no way of enforcing this without using
        supplementary tables or views./n  Row-level filtering labels each row with a
        particular security categorisation (not visible to the user) reminiscent of
        the Bell-LaPadula Model.  Thus when a user SELECTs data from a table, the
        rows returned are not the full result set demanded by the SQL that was
        executed, but a subset of the result set that is compliant with the row-level
        filtering policy and the user's access permissions.",
      "Threats": {
        "Confidentiality": "Disclosure of material to users not authorised to view
          the material"
      },
      "Best Practices": "Row-level filtering should be enabled wherever there is a
        requirement for different users, roles or groups to view subsets of data
        inside a single relational table.  If views are currently used to achieve the
        same effect, consider replacing this with row-level filtering for more
        granular control.  Note that controlling row-level filtering policy can be
        more administratively demanding than managing view permissions, so have this
        in mind before deciding to implement this control.  Row-level filtering is
        not available in all RDBMS platforms.  It is available in Microsoft SQL
        Server version 2016 and above, and Azure DB, Oracle BI Enterprise 12c and
        above, and not currently available in MySQL.",
      "Further Reading": {
        "Link 1": "https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/sql
          /granting-row-level-permissions-in-sql-server",
        "Link 2": "https://docs.oracle.com/middleware/12211/biee/BIEMG/GUID-1FDC0A15
          -9DE7-4838-9C0E-03290F5558B2.htm#dataaccess_row"
      },
      "Related Topics": {
        "1": {
          "Abstract Concepts": "Bell-LaPadula Model"
        },
        "2": {
          "Access Control": "Mandatory Access Control"
        },
        "3": {
          "Confidentiality": "Negative Authorisation"
        },
        "4": {
          "Confidentiality": "View-Based Access"
        },
        "5": {
          "Data Categorisation": "Data Labelling"
        },
        "6": {
          "Obfuscation": "Data Masking"
        }
      }
    }
  }
}
```

Figure 5.2: JSON document detailing a topic record

5.1.3.3 Windows Help

For the third, and more practical, implementation the data was put into a hierarchical form using the tool HelpNDoc (IBE Software, available at https://www.helpndoc.com). This involved creating the content categories, header and footer templates and populating the pages according to the DSF design template (identifier, title, precis and so on). The following screenshots demonstrate the final product. Annotations in **bold red text** have been added to the figures for illustration only. All the content derives from the research outcomes (insights, conjectures and researched information) of the primary survey research and the secondary literature review as a consequence of the analysis described in Chapter 4:

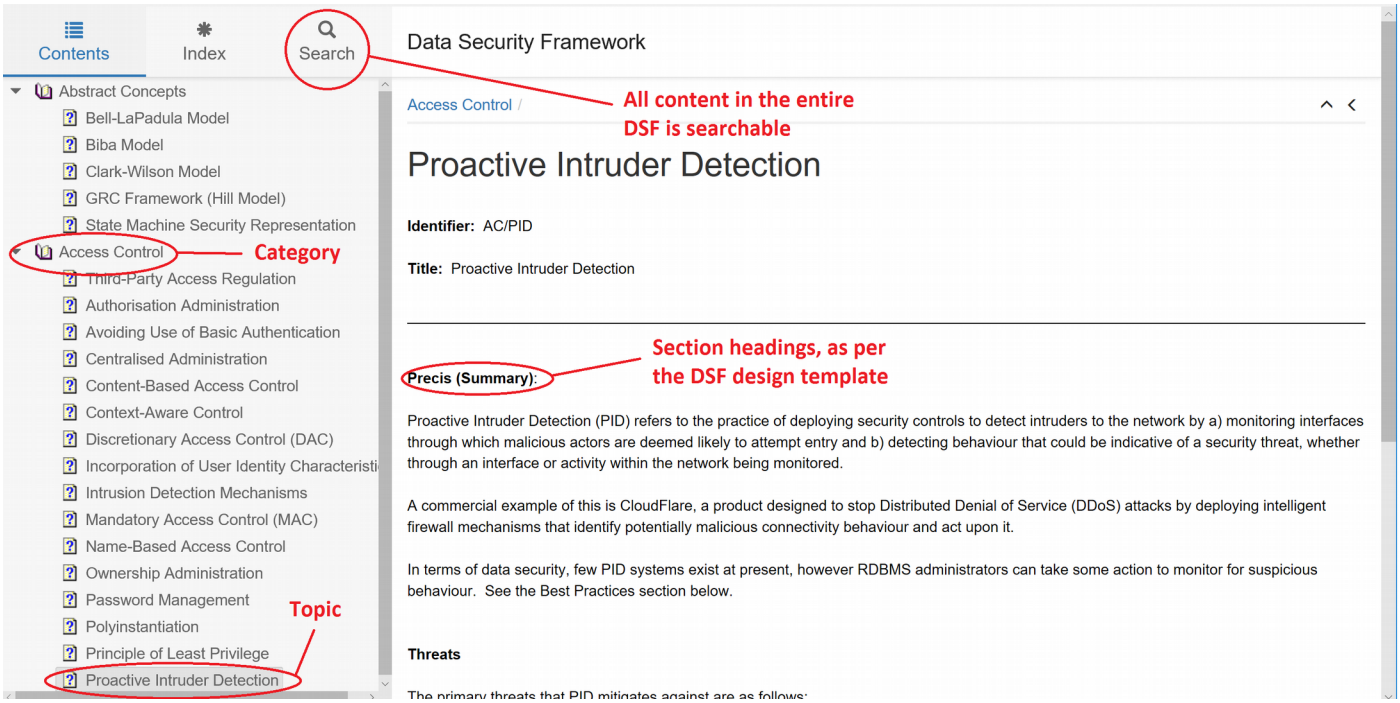


Figure 5.3(a): Screenshot (1) of the DSF, implemented for Windows Help

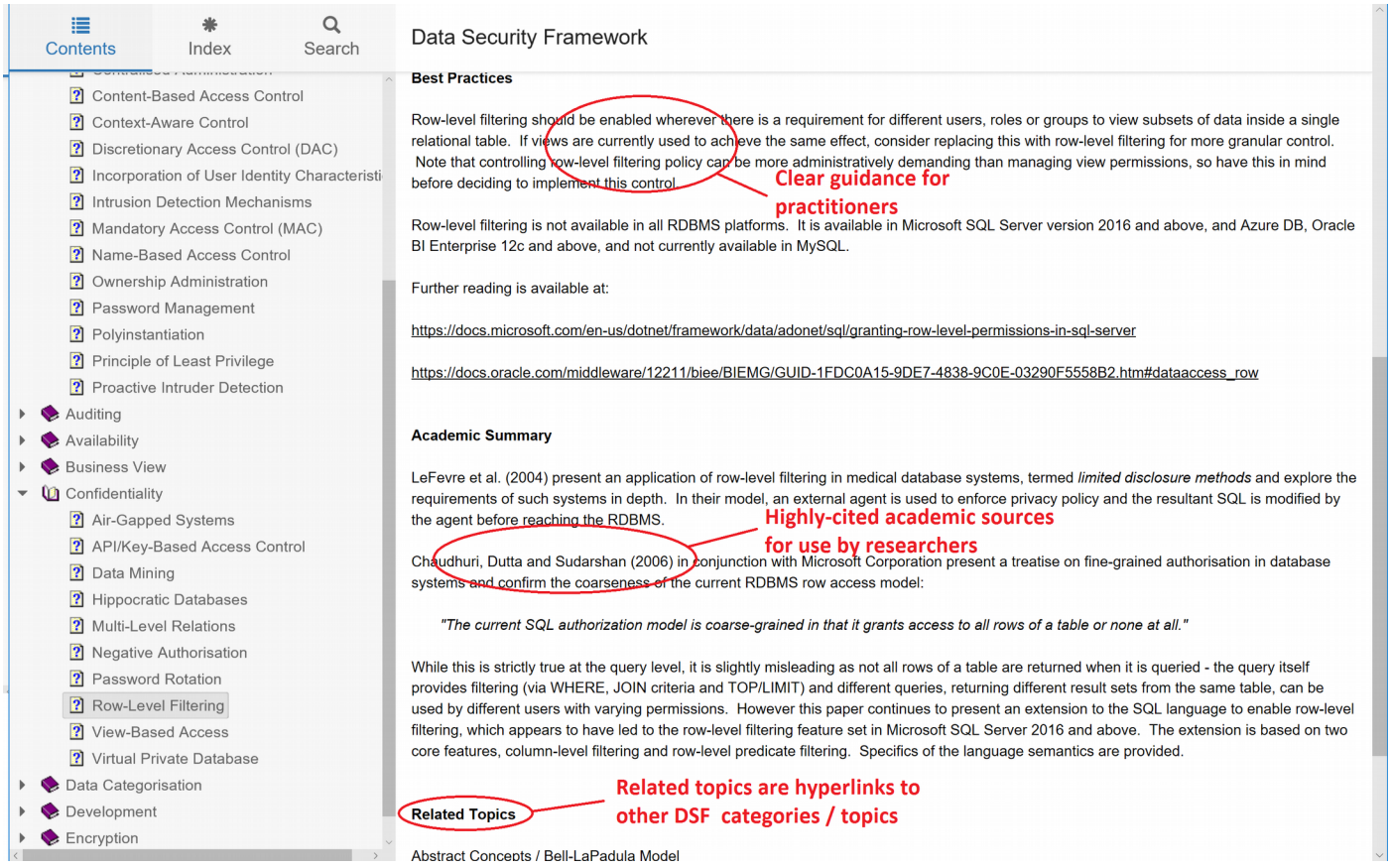


Figure 5.3(b): Screenshot (2) of the DSF, implemented for Windows Help

These sample implementations show that the Data Security Framework is viable and can be created as a tool. However, some further work would be needed to create a process to manage and curate the content so that it

remained accurate. While the DSF as shown represents the possibility of a complete dictionary, further enhancements are possible – the cross- or sub-categorisation of topics into current vs. future practices, for example, or supplementary training material links. Further research, such as a longitudinal study or case study to assess the practical usefulness of the tool to an organisation, would be valuable.

5.1.4 Objective 4: GAP Analysis

The following two figures show an aggregated SWOT view of each domain, industrial and academic, forming the conclusions of the GAP analysis. To reiterate, these diagrams help show gaps where industry could be implementing techniques born of research and where research is necessary to develop techniques pioneered in industry. These conclusions come from the data presented in Section 4.3.3, Appendix F and Appendix G.

Industrial

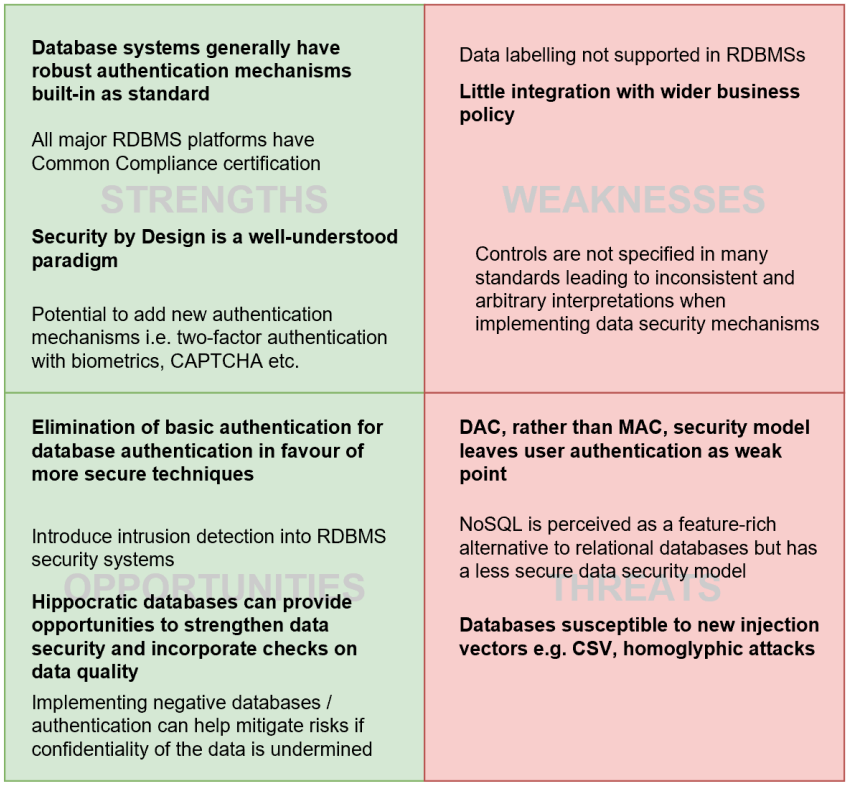


Figure 5.4: SWOT analysis for industrial applications of data security techniques

Academic

Figure 5.5: SWOT analysis of academic research opportunities

5.1.5 Objective 5: Validation and Reflection

This objective was to consider the outcomes of the research given external technological and cultural factors. Although not a primary aim of the research, the outputs did yield insights, for example into the attitudes of some practitioners towards integration with IT governance standards as discussed in Section 4.3.5.

The main conclusion related to this objective is that technological development is progressing at a pace that challenges the speed at which research and governance developments can keep up. This is evidenced particularly in cloud computing, with platform-as-a-service data security challenges not reflected in current best practice nor supported wholly by a significant body of academic research output.

5.2 Conclusions about the research aim

5.2.1 Effectiveness of the Aim

The aim was a consolidated restatement of the research objectives, which were primarily to produce two deliverables; a Data Security Framework for use by practitioners and researchers, and a GAP analysis to determine which areas of data security had opportunities for implementation and further research. Whilst the aim was effective in defining the deliverables, some areas were overlooked, for example the applicability of the DSF across

different industries, dissemination of the DSF to industry and academia, validation of the results and ensuring continual updates to keep the content refreshed. On reflection, the aim should have encompassed these points and also some further research to measure the value of the outputs with the intended audience.

However in general the aim was stated clearly and in such a way that the deliverables could be defined and met, and the outcomes add some structure to an area of computing which is ill-defined to the benefit of those working in the area. From this perspective, the aim was sufficient to determine the course of the research.

5.2.2 General Conclusions

The following bullet points summarise the conclusions of the research as a whole:

- The persistent evidence of data breaches that affect organisations show that data security remains a current concern;
- Current information security standards encompass data security but do not provide for details on the different aspects, interpretations, implementations or research questions surrounding the area;
- Practitioners use similar best practices derived largely from shared knowledge and practitioner literature, but lack a unified security framework;
- Research material is broad and deep when considering long-standing information security concepts such as access control, but less so for newer technologies and in some areas, research is not keeping up with industry developments;
- The Data Security Framework produced provides a unified reference model for practitioners and researchers to determine the best course of action in applying security controls to their organisations or scoping out new areas for research, respectively;
- The content of the Data Security Framework expressed in this research (categories and topics) is a starting point for the further development of the DSF to be suited to wider industrial and academic areas;
- The GAP analysis provided is a summary of at least some of the areas of data security that have differing industry and academic research coverage;
- The pace of technological development is causing serious issues in ensuring good data security throughout the design and development lifecycle;
- More validation is required to make sure that the DSF is a viable and useful tool for users.

5.3 Further work

5.3.1 Suggestions for further work

Further work in data security research is recommended. This should include a more systematic literature review to encompass the current authentication mechanisms and security features of all major relational and non-relational database management systems. This would have the benefit of identifying weaknesses where major themes and best practices in data security are not followed or implemented correctly and provide opportunities for practitioners to shore up defences against potential threats.

Further validation is required that the Framework is fit for purpose, and the GAP analysis could be expanded to use other analysis techniques combined with further interviews or other engagement with the academic domain to provide more detail and avenues for investigation.

5.4 Implications of the research

In the best case, where the Data Security Framework and GAP analysis continues to be expanded and developed, and dissemination to appropriate users is achieved, the following list of implications is anticipated:

- Convergence of best practices between practitioners in data security operating in different areas of industry;
- Movement towards a dedicated data security IT governance framework that allows for the specific consideration of data security (rather than general information technology) level issues;
- Knowledge sharing of the DSF as a dictionary of recommendations and techniques for direct application as data security controls;
- Knowledge sharing of the DSF and GAP analysis as a central reference point for researchers looking for future directions in data security matters.

5.5 Reflection on the experience of the research process

I have enjoyed conducting this research and have learned much, especially in planning, conducting and managing a complex research project. I encountered several difficulties which needed to be overcome: for example, how to manage a large amount of unstructured data collected narratively and stored on many different mediums; how to find the most appropriate ways of turning this data into actionable insights; dealing with non-empirical data (which challenged me, as my prior experience is in analysing scientific, quantitative data); and time management, as events occurred which forced me to alter my progress plan.

If I had the opportunity to conduct this research afresh I would have limited the scope further, to ensure that a clear path could be drawn between collected data and conclusions, which I am aware is difficult to follow with the current

wide scope; I would also have included some method of external validation i.e. external review and feedback via interviews, rather than relying solely on internal consistency, and engaged more fully with the academic population.

However, I am happy with the results of this research and am confident that the deliverables add to the current body of knowledge, even if this contribution is relatively minor. I have more awareness of how such research should be carried out and the feedback received during T802 has been invaluable in helping to refine my research skills.

5.6 Chapter Summary

In this chapter we used the preliminary conclusions drawn in Chapter 4 to reach firm conclusions and presented three different implementations of the Data Security Framework. We showed how SWOT analysis could be used to summarise the results of the GAP analysis to highlight areas for future work by both industry and academic experts. We reflected on the implications of the research and the research process.

6. References

Abrahamsson, P., Salo, O., Ronkainen, J. and Warsta, J. (2002) 'Agile software development methods: Review and analysis', *VTT Publication 478*, pp. 107 [Online]. Available at: http://www.vtt.fi/inf/pdf/publications/2002/P478.pdf (Accessed 27 Dec 2017).

Afyouni, H.A. (2006). *Database Security and Auditing: Protecting Data Integrity and Accessibility.* Canada, Thomson Course Technology.

Arlitsch, K. and Edelman, A. (2014) 'Staying safe: Cyber security for people and organisations', *Journal of Library Administration*, vol. 54, pp. 46-56 [Online]. Available at: http://libezproxy.open.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=95326375&site=ehost-live&scope=site (Accessed 14 Oct 2017).

Astrahan, M.M., Blasgen, M.W., Chamberlin, D.D., Eswaran, K.P., Gray, J.N., Griffiths, P.P., King, W.F., Lorie, R.A., McJones, P.R., Mehl, J.W. and Putzolu, G.R. (1976) 'System R: Relational approach to database management', *ACM Transactions on Database Systems (TODS)*, vol 1, no. 2, pp.97-137.

Atzeni, P., Jensen, C.S., Orsi, G., Ram, S., Tanca, L. and Torlone, R. (2013) 'The relational model is dead, SQL is dead, and I don't feel so good myself'. *ACM SIGMOD Record*, vol 42, no. 2, pp.64-68 [Online]. Available at: http://torlone.dia.uniroma3.it/pubs/sr2013.pdf (Accessed 09 Nov 17)

Australian Financial Review (2014) 'Telstra fined over data breach', 3 October [Online]. Available at: http://www.afr.com/technology/telstra-fined-over-data-breach-20140310-ixlus (Accessed 15 Jan 2018)

Bamrara, A. (2015) 'Evaluating Database Security and Cyber Attacks: A Relational Approach', *Journal of Internet Banking and Commerce*, vol. 20, no. 2, pp. 1-17 [Online]. Available at: http://pmt-eu.hosted.exlibrisgroup.com/44OPN_VU1:EVERYTHING:TN_proquest1799378132 (Accessed 14 Oct 2017)

Basta, A. and Zgola, M. (2011) *Database Security*. Cengage Delmar Learning.

Beauchemin, B. (2012) 'SQL Server 2012 Security Best Practices – Operational and Administrative Tasks', *Microsoft Corporation (White Paper)*, pp. 25-27 [Online]. Available at: http://download.microsoft.com/download/8/F/A/8FABACD7-803E-40FC-ADF8-355E7D218F4C/SQL_Server_2012_Security_Best_Practice_Whitepaper_Apr2012.docx (Accessed 28 Nov 17).

Bell, D.E. and LaPadula, L.J. (1976) 'Secure computer system: unified exposition and Multics interpretation' [Online]. Available at: https://csrc.nist.gov/csrc/media/publications/conference-paper/1998/10/08/proceedings-of-the-21st-nissc-1998/documents/early-cs-papers/bell76.pdf (Accessed 10 Mar 2018).

Bentley, C. (2012). *PRINCE2: A Practical Handbook.* London : Routledge.

Bertino, E. and Sandhu, R. (2005) 'Database security - concepts, approaches, and challenges', *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 1, pp.2-19 [Online]. Available at: https://search.proquest.com/openview/e03943036a308be889aa30aece031fe3/1?pq-origsite=gscholar&cbl=27603 (Accessed 24/10/2017).

Boehmer, W. (2008) 'Appraisal of the effectiveness and efficiency of an information security management system based on ISO 27001', *Second International Conference on Emerging Security Information, Systems and Technologies (SECURWARE'08)*, pp. 224-231 [Online].

Available at: http://ieeexplore.ieee.org/abstract/document/4622587/?reload=true (Accessed 06 Apr 18).

Boyatzis, R.E. (1998). *Transforming qualitative information: Thematic analysis and code development*. London, Sage.

Boyd, D. and Crawford, K. (2012) 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon', *Information, Communication & Society*, vol. 15, no. 5, pp.662-679 [Online]. Available at: http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878 (Accessed 10 Jan 18).

Braun, V., Clarke, V. and Terry, G. (2012). 'Thematic Analysis'. APA Handbook of Research Methods in Psychology, Vol. 2, pp.57-71.

British Broadcasting Corporation (2013). 'Adobe hack: At least 38 million accounts breached' [Online]. Available at: https://www.bbc.com/news/technology-24740873 (Accessed 26 Aug 18).

Cachin, C. and Schunter, M. (2011) 'A cloud you can trust', *IEEE Spectrum*, vol. 48, no. 12, pp.28-51.

Calder, A. and Watkins, S. (2005). *IT Governance: A Manager's Guide to Data Security and BS7799/ISO 17799*, 3rd edition. London, Kogan-Page.

Carter, P. (2016). *Securing SQL Server*. New York, Apress.

Codd, E.F. (1969) 'A relational model of data for large shared data banks', *Communications of the ACM*, vol. 13, no. 6, pp.377-387.

Codd, E.F. (1990). *The relational model for database management: version 2*. Boston, Addison-Wesley Longman Publishing Co., Inc.

Common Criteria (2018) 'CC v3.1 Release 5' [Online]. Available at: https://www.commoncriteriaportal.org/cc/ (Accessed 07 Apr 18).

Commons, House of. Great Britain. Parliament, (1998). *Data Protection Act*. London, The Stationery Office [Online]. Available at: http://www.legislation.gov.uk/ukpga/1998/29/contents (Accessed 28 Nov 17)

Connolly, T. M. and Begg, C. E. (2004). *Database Systems: A Practical Approach to Design, Implementation and Management*, ch. 13, pp. 408-413. New York, Pearson Education.

Coronel, C. and Morris, S. (2016). *Database systems: design, implementation and management.* Boston, Cengage Learning.

Creswell, J.W. and Plano-Clark, V.L. (2007). *Designing and conducting mixed methods research*. California, Sage Publications.

Custer, W.L. (2010), 'Information Security Issues in Higher Education and Institutional Research', *New Directions for Institutional Research*, vol. 146, pp. 23-50 [Online]. Available at: http://libezproxy.open.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=52470905&site=ehost-live&scope=site (Accessed 14 Oct 2017)

Damele, B. and Stampar, M. (2018). 'SQLMAP' [online]. Available at: http://sqlmap.org (Accessed 9 Jun 18).

Date, C.J. (1986). *Relational database: selected writings (vol. 1)*. Massachusetts, Addison-Wesley.

de Vaus, D. ed. Greenfield, T. (2002). *Research Methods for Postgraduates*, ch. 22, pp. 172-182. Hodder Education, London.

Docker Incorporated (2017) 'Dockerize PostgreSQL' [Online]. Available at: https://docs.docker.com/engine/examples/postgresql_service/ (Accessed 22 Jan 2018)

European Union, Regulation (EU) 2016/679 (General Data Protection Regulation), 2016.

Gaetjen, S., Knox, D. and Maroulis, W. (2015). *Oracle Database 12c Security*. New York, McGraw-Hill Education.

Ge, X., Polack, F. and Laleau, R. (2004) 'Secure databases: an analysis of Clark-Wilson [sic] model in a database environment', *International Conference on Advanced Information Systems Engineering*, pp. 234-247 [Online]. Available at: https://www.researchgate.net/profile/Fiona_Polack/publication/220920890_Secure_Databases_An_Analysis_of_Clark-Wilson_Model_in_a_Database_Environment/links/0c96051a4f0aa75c7f000000.pdf (Accessed 24 Jan 2018).

Gick, C. and Richins, J. (2008) 'Engine Separation of Duties for the Application Developer' [Online]. Available at: https://technet.microsoft.com/en-us/library/cc974525(v=sql.100).aspx (Accessed 28 Nov 17).

Glaser, B. and Strauss, A.L. (1967). *The discovery of grounded theory: strategies for qualitative research*. Chicago, Aldine.

Greenfield, T. (2002) (ed). *Research Methods for Postgraduates*. London, Hodder Education.

Gressin, S. and Charleston, S. (2017) 'The Equifax Data Breach: What to Do' [Online]. Available at: https://www.netcreditunion.com/equifaxdatabreach/ (Accessed 15 Nov 17).

Gubrium, J.F. and Holstein, J.A. eds. (2002). *Handbook of interview research: Context and method*. London, Sage.

Gupta, K., Malik, A., Pawar, S., and Patil, J. (2016) 'Database Security Two Way Authentication Using Graphical Password', *International Journal of Engineering Research and Applications*, vol. 6, no. 4, pp. 100-103 [Online]. Available at: http://pmt-eu.hosted.exlibrisgroup.com/44OPN_VU1:EVERYTHING:TN_doaj_soai_doaj_org_article_b3254b715d044ac28c24415bc07d9785 (Accessed 14 Oct.2017).

Harter, D.E., Krishnan, M.S. and Slaughter, S.A. (2000) 'Effects of process maturity on quality, cycle time, and effort in software product development', *Management Science*, vol. 46, no. 4, pp.451-466.

Hill, D.G. (2009). *Data Protection: Governance, Risk Management and Compliance*. Florida, Auerbach Publications.

Hove, S.E. and Anda, B. (2005). 'Experiences from conducting semi-structured interviews in empirical software engineering research'. *11th IEEE International Symposium in Software Metrics,* pp. 10-13.

Hubermann, A. and Miles, M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. London, Sage.

Huygh, T., De Haes, S., Joshi, A. and Van Grembergen, W. (2018) 'Answering Key Global IT Management Concerns Through IT Governance and Management Processes: A COBIT 5 View'

[Online]. Available at: https://repository.uantwerpen.be/docman/irua/481d07/147825.pdf (Accessed 02 Mar 2018).

International Standards Organisation (2013) 'ISO/IEC 27001:2013' [Online]. Available at: https://www.iso.org/standard/54534.html (Accessed 28 Nov 17).

International Standards Organisation (2016) 'ISO/IEC 9075-1:2016 - Information technology - Database languages - SQL - Part 1: Framework (SQL/Framework)' [Online]. Available at: https://www.iso.org/standard/63555.html (Accessed 5 Oct. 2017).

Invesp. Inc (2018) cites GlobalWebIndex (2015). 'Social Media Engagement' [online]. Available at: https://www.invespcro.com/blog/social-media-engagement/ (Accessed 01 Jun 18).

ISACA (2017) 'COBIT 5' [Online]. Available at: http://www.isaca.org/COBIT/Pages/default.aspx (Accessed 28 Nov 17).

Jajodia, S. (1996) 'Database security and privacy', *ACM Computing Surveys (CSUR)*, vol. 28, no. 1, pp.129-131.

Kettle, J. and Context Information Security (2014) 'Comma Separated Vulnerabilities' [Online]. Available at: https://www.contextis.com/blog/comma-separated-vulnerabilities (Accessed 28 Nov 17).

Kindy, D.A. and Pathan, A.S.K. (2011) 'A survey on SQL injection: Vulnerabilities, attacks, and prevention techniques', *IEEE 15th International Symposium on Consumer Electronics (ISCE)*, pp. 468-471 [Online]. Available at: http://irep.iium.edu.my/769/1/ISCE2011_paper323.pdf (Accessed 09 Nov 2017).

Kirti, G., Gupta, R., Biswas, K. and Turlapati, R.R.S., Oracle International Corp. (2017). 'Techniques for cloud security monitoring and threat intelligence'. U.S. Patent 9,692,789.

Klahr, R., Shah, J.N., Sheriffs, P., Rossington, T., Pestell, G., Button, M., Wang, V., Department for Culture, Media and Sport, IPSOS Mori, and University of Portsmouth (2017) 'Cyber Security Breaches Survey 2017', [online] Available at: https://www.gov.uk/government/uploads/system/uploads/attaHTMent_data/file/609186/Cyber_Security_Breaches_Survey_2017_main_report_PUBLIC.pdf (Accessed 02 Nov 2017).

Landwehr, C. (1991) 'Formal Models for Computer Security', *ACM Computing Surveys*, vol. 13, no. 3 [Online]. Available at: http://crypto.stanford.edu/~ninghui/courses/Fall03/papers/landwehr_survey.pdf (Accessed 18 Nov 17).

Lane, V.P. and Macmillan Education UK (1985) 'The Security Role of Components of Computer Configurations', *Security of Computer Based Information Systems*, pp. 54-74.

Layton, R. and Watters, P.A. (2014), 'A methodology for estimating the tangible cost of data breaches', *Journal of Information Security and Applications*, vol. 19, no. 6, pp.321-330 [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S2214212614001483 (Accessed 20 Jan 2018).

Learned, E. P. (1969). *Business Policy: Text and Cases.* Chicago: Richard D. Irwin.

Lewis, P.M., Bernstein, A.J. and Kifer, M. (2002). *Databases and transaction processing: an application-oriented approach*. Reading, Addison-Wesley.

Lu, H., Chan, H.C. and Wei, K.K. (1993) 'A Survey on Usage of SQL', *ACM SIGMOD Record*, vol. 22, no. 4, pp.60-65.

Machanavajjhala, A. and Reiter, J. P. (2012). 'Big privacy: protecting confidentiality in big data.' *XRDS: Crossroads, the ACM Magazine for Students - Big Data*, vol. 19, no. 1, pp.20-23.

Mansfield-Devine, S. (2015) 'The Ashley Madison affair', Network Security, vol. 9, pp.8-16 [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S1353485815300805 (Accessed 10 Jan 2018).

Mantelero, A. (2013) 'The EU Proposal for a General Data Protection Regulation and the roots of the 'right to be forgotten'', *Computer Law & Security Review*, vol. 29, no. 3, pp.229-235 [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S0267364913000654 (Accessed 14 Feb 2018).

McCumber, J. (1991), 'Information systems security: A comprehensive model', *Proceedings of the 14th National Computer Security Conference, National Institute of Standards and Technology* [Online]. Available at: http://trygstad.rice.iit.edu:8000/Government%20Documents/NSTISS/NSTISSI4011Annex.rtf (Accessed 18 Feb 2018).

Microsoft Corporation (2005), 'Data Confidentiality' [Online]. Available at: https://msdn.microsoft.com/en-us/library/ff650720.aspx (Accessed 17 Nov 17).

Microsoft Corporation (2005b) 'Data Integrity' [Online]. Available at: https://technet.microsoft.com/en-us/library/ms184276(v=sql.105).aspx (Accessed 17 Nov 17).

Microsoft Corporation (2012) 'SQL Server 2012 Security Best Practices – Operational and Administrative Tasks' [Online]. Available at: http://download.microsoft.com/download/8/f/a/8fabacd7-803e-40fc-adf8-355e7d218f4c/sql_server_2012_security_best_practice_whitepaper_apr2012.docx (Accessed 5 Oct 2017).

Microsoft Corporation (2016) 'Stretch Database' [Online]. Available at: https://docs.microsoft.com/en-us/sql/sql-server/stretch-database/stretch-database (Accessed 21 Jan 2018).

Microsoft Corporation (2016b) 'FIPS 140-2 Compliance' [Online]. Available at: https://technet.microsoft.com/en-us/library/bb326611(v=sql.105).aspx (Accessed 07 Apr 18).

Microsoft Corporation (2017). 'Service Master Key' [Online]. Available at: https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/service-master-key (Accessed 02 Nov 2017).

Mitre Corporation (1977). *Integrity Considerations for Secure Computer Systems*. Washington, Defence Technical Information Center (DTIC), US Department of Defence.

National Institute of Standards and Technology (NIST) (2013) 'NIST Special Publication 800-53 Revision 4: Security and Privacy Controls for Federal Information Systems and Organisations' [Online]. Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf (Accessed 04 Apr 2018).

Neti, S. (2011). 'Social Media and its Role in Marketing', *International Journal of Enterprise Computing and Business Systems*, vol 1., issue 2, pp. 1-15.

Neumann, P.G. (1993) 'Security criteria for electronic voting', *16th National Computer Security Conference*, vol. 29 [Online]. Available at: http://www.csl.sri.com/~neumann/ncs93.html (Accessed 02 Jan 2018).

Offensive Security (2018). 'Linux Kali' [online]. Available at: https://www.kali.org (Accessed 9 Jun 18).

Office of the Australian Information, Commissioner (2012), 'Telstra breaches Privacy Act' [Online]. Available at: https://www.oaic.gov.au/media-and-speeches/media-releases/telstra-breaches-privacy-act (Accessed 15 Jan 2018).

Okman, L., Gal-Oz, N., Gonen, Y., Gudes, E. and Abramov, J. (2011) 'Security issues in nosql databases', *IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications* (TrustCom), pp. 541-547 [Online]. Available at: http://ieeexplore.ieee.org/abstract/document/6120863/ (Accessed 15 Jan 2018).

Olivier, M.S. (2002) 'Database privacy: balancing confidentiality, integrity and availability', *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 2, pp.20-27 [Online]. Available at: http://mo.co.za/open/dbpriv.pdf (Accessed 25 Oct 2017).

Oracle Corporation (2017) 'Adaptive Intelligent Apps' [Online]. Available at: https://cloud.oracle.com/en_US/adaptive-intelligent-apps (Accessed 21 Jan 2018).

Oracle Corporation (2017) 'Maximise Availability with Oracle Database 12c Release 2 (white paper)' [Online]. Available at: www.oracle.com/technetwork/database/availability/maximum-availability-wp-12c-1896116.pdf (Accessed 17 Nov 17).

Oracle Corporation (2018) 'Oracle Security Evaluations' [Online]. Available at: https://www.oracle.com/technetwork/topics/security/security-evaluations-099357.html (Accessed 07 Apr 18).

Oxford University Press (2018) 'Oxford Dictionaries (English): Definition of "schema"' [Online]. Available at: https://en.oxforddictionaries.com/definition/schema (Accessed 30 Aug 18).

PCI Security Standards Council (2017) 'PCI DSS v3.2' [Online]. Available at: https://www.pcisecuritystandards.org/document_library?category=pcidss&document=pci_dss (Accessed 28 Nov 17).

Pernul, G. (1994) 'Database security', *Advances in Computers*, vol. 38, pp.1-72.

Qiu, B., Fang, N. and Wenyin, L. (2010). 'Detect visual spoofing in Unicode-based text', *Pattern Recognition (ICPR), 20th International Conference*, pp. 1949-1952 [Online]. Available at: http://ieeexplore.ieee.org/abstract/document/5597240/ (Accessed 25 Oct 2017).

Rask, A., Rubin, D. and Neumann, B. (2005) 'Implementing Row- and Cell-Level Security in Classified Databases Using SQL Server 2005' [Online]. Available at: https://technet.microsoft.com/en-us/library/cc966395.aspx (Accessed 28 Nov 17).

Ratha, N. K., Connell, J. H. and Bolle, R. M. (2001) 'Enhancing security and privacy in biometrics-based authentication systems.' *IBM Systems Journal*, vol. 40, no. 3, pp. 614-634 [Online]. Available at: https://ieeexplore.ieee.org/abstract/document/5386935/ (Accessed 20 Aug 18).

Rindell, K., Hyrynsalmi, S. and Leppänen, V. (2015) 'A comparison of security assurance support of Agile software development methods', *ACM Proceedings of the 16th International Conference on Computer Systems and Technologies*, pp. 61-68 [Online]. Available at: https://dl.acm.org/citation.cfm?id=2812431 (Accessed 21 Jan 2018).

Romanosky, S., Hoffman, D. and Acquisti, A. (2014) 'Empirical analysis of data breach litigation', *Journal of Empirical Legal Studies*, vol. 11, no. 1, pp.74-104 [Online]. Available at: http://onlinelibrary.wiley.com/doi/10.1111/jels.12035/full (Accessed 28 Nov 2017).

Roshanbin, N. and Miller, J. (2011) 'Finding homoglyphs - a step towards detecting unicode-based visual spoofing attacks', *Web Information System Engineering (WISE)*, pp.1-14 [Online].

Available at: https://link.springer.com/10.1007%2F978-3-642-24434-6_1 (Accessed 20 Feb 2018).

Saltzer, J.H. and Schroeder, M.D. (1975) 'The protection of information in computer systems', *Proceedings of the IEEE*, vol. 63, no. 9, pp.1278-1308 [Online]. Available at: http://ieeexplore.ieee.org/abstract/document/1451869/ (Accessed 10 Jan 2018).

Saunders, M.L. and Lewis, P. (2009). *Research Methods for Business Students*, pp. 123-124, 167. New Jersey, Pearson.

Schneider, F.B. (2003), 'Least privilege and more', *IEEE Security & Privacy*, vol. 99, no. 5, pp.55-59 [Online]. Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1236236 (Accessed 09 Nov 2017).

Sharma, N.K. and Dash, P.K. (2012), 'Effectiveness of ISO 27001 as an information security management system: An analytical study of financial aspects', *Far East Journal of Psychology and Business*, vol. 9, no. 3, pp.42-55 [Online]. Available at: http://www.academia.edu/download/32005263/effectivenes_of_iso_27001.pdf (Accessed 08 Apr 2018).

Shete, S.S. and Kulakrni, C.S. (2015) 'Role-Based Access Control within RDBMS', *International Journal of Advanced Research in Computer Science*, vol. 6, no. 7 [Online]. Available at: http://www.ijarcs.info/index.php/Ijarcs/article/viewFile/2584/2572 (Accessed 09 Nov 2017).

Sikeridis, D., Papapanagiotou, I., Rimal, B.P. and Devetsikiotis, M. (2017) 'A Comparative Taxonomy and Survey of Public Cloud Infrastructure Vendors', *Cornell University Library (arxiv.org)* [Online]. Available at: https://arxiv.org/pdf/1710.01476 (Accessed 10 Nov 17).

Solid IT Consulting and Software Development gmbh (2018). 'Knowledge Base of Relational and NoSQL Database Management Systems: DB Engines Ranking' [Online]. Available at: https://db-engines.com/en/ranking (Accessed 28 Aug 18).

Snegovaya, M., (2015). 'Putin's information warfare in Ukraine. Soviet Origins Of Russia's Hybrid Warfare'. *The Russia Report*, Washington.

Stallings, W. and Brown, L. (2012). *Computer Security Principles and Practice*. New Jersey, Pearson.

Stonebraker, M. (1986). *The INGRES papers: anatomy of a relational database system*. Boston, Addison-Wesley Longman Publishing Co., Inc.

Sugimori, Y., Kusunoki, K., Cho, F. and Uchikawa, S. (1977). 'Toyota production system and Kanban system materialization of just-in-time and respect-for-human system', *International Journal of Production Research*, vol. 15, no. 6, pp. 553-564 [Online]. Available at: https://www.tandfonline.com/doi/pdf/10.1080/00207547708943149 (Accessed 21 Jul 2018).

Tchernykh, A., Schwiegelsohn, U., Talbi, E.G. and Babenko, M. (2016) 'Towards understanding uncertainty in cloud computing with risks of confidentiality, integrity, and availability', *Journal of Computational Science* [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S1877750316303878 (Accessed 25/10/2017).

Telstra (2011), Public Announcement, 'Telstra Exchange'. Available at: https://exchange.telstra.com.au/public-announcement/ (Accessed 15 Jan 2018).

The Register (2012) 'Hackers expose 6.5 MILLION 'LinkedIn passwords'', 6 June [Online]. Available at: http://www.theregister.co.uk/2012/06/06/linkedin_password_leak/ (Accessed 18 Jan 2018).

Thomas, K., Li, F., Zand, A., Barrett, J., Ranieri, J., Invernizzi, L., Markov, Y., Comanescu, O., Eranti, V., Moscicki, A. and Margolis, D. (2017). 'Data Breaches, Phishing, or Malware?: Understanding the Risks of Stolen Credentials.' *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, October, pp. 1421-1434. ACM.

Thusoo, A., Shao, Z., Anthony, S., Borthakur, D., Jain, N., Sen Sarma, J., Murthy, R. and Liu, H. (2010), 'Data warehousing and analytics infrastructure at Facebook', *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, pp. 1013-1020 [Online]. Available at: https://dl.acm.org/citation.cfm?id=1807278 (Accessed 10 Jan 2018).

Townsend, W.W. (1924). *Sales Bondmanship*, pp. 109. New York: Henry Holt.

Under Armour Performance Inc. (2018). 'Press release: Under Armour Notifies MyFitnessPal Users Of Data Security Issue' [Online]. Available at: http://investor.underarmour.com/news-releases/news-release-details/under-armour-notifies-myfitnesspal-users-data-security-issue (Accessed 28 Aug 18).

Vavilis, S., Egner, A., Petkovic, M., Zannone, N. (2015), 'An anomaly analysis framework for database systems', *Computers & Security*, vol. 53, pp. 156-173 [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S0167404815000905 (Accessed 14 Oct 2017).

Woo, V. and Exploit Database (2017), 'Apache Struts 2.3.5 < 2.3.31 / 2.5 < 2.5.10 – Remote Code Execution' [Online]. Available at: https://www.exploit-db.com/exploits/41570/ (Accessed 5 Oct 2017).

Yasnoff, W.A. (2016) 'Breach Risk Magnitude: A Quantitative Measure of Database Security', *AMIA Annual Symposium Proceedings*, pp. 1258–1263 [Online]. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333333/pdf/2499885.pdf (Accessed 14 Oct 2017).

7. Extended Abstract

(1,231 words, 5 pages)

Introduction

Relational Database Management Systems (RDBMSs) are software systems concerned with the management and administration of data. Based on set theory, data is held in relations, which are implemented as tables. Each table has rows and columns where each row contains a set of corresponding facts and each column is an attribute of that set, describing a fact.

RDBMSs are used worldwide to store and manage data, from small single-table instances to multi-petabyte distributed systems. Of all database systems available, 4 out of the top 5 by usage are relational databases (Solid IT Consulting and Software Development gmbh, 2018). The job of an RDBMS is to store data *confidentially*, with *integrity* and preserving *availability* – collectively known as the C.I.A. principles (Olivier, 2002). We will refer to these concepts using a single umbrella term, *data security*. The following diagram shows where RDBMSs fit into the application stack:

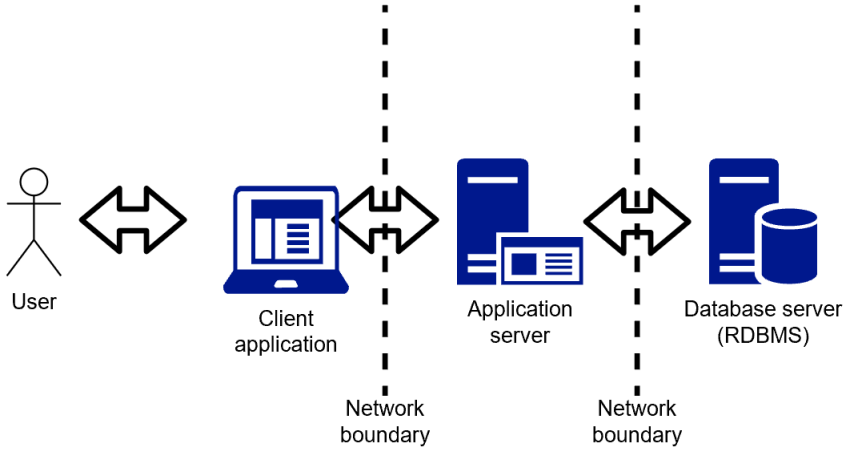


Figure 1: Basic interactions between the user and the database server

Aim, Objectives and Justification for the Research

Currently, information security governance is mandated by standards which encompass many different areas of information technology, including network security, user administration, production of policies and procedures, hardware and physical security, application security and data security. As such there is no set of dedicated standards for ensuring security at the database level. As RDBMS systems operate on very different paradigms to applications and indeed the rest of the stack, special consideration is required when planning a strategy for implementing data security and when choosing which security controls are most appropriate for the circumstance. Database systems

are also frequently under attack and are a valuable target for malicious actors. High-profile data breaches have negative repercussions on organisations, from reputational damage to financial loss and even to the inability of the organisation to continue operations.

Unfortunately, there does not exist any comprehensive source of information on design and deployment of data security for the practitioner. Nor, from a research perspective, is there a central guide to the various aspects of data security that can serve as an up-to-date summary of the state of data security research. Instead, governance standards stop short of recommending specific controls and much is left up to the individual practitioner. Research into the area is also incomplete, with viable, well-developed ideas lacking implementation, and *vice versa*, many developments in industry lacking a research pedigree.

The central objective of this research is to investigate and produce a Data Security Framework (DSF) that will both a) enable practitioners to refer to a single reference guide to all matters relating to data security and b) provide guidance for researchers in the field to understand the current state of knowledge and provide guidance on future research opportunities.

Methodology

The research is split into three channels – literature review, primary research (questionnaire) and primary research (interviews). Along each channel, data collection, data analysis and the production of conclusions are undertaken. This is a mixed-methods study drawing inspiration from the pragmatic philosophical branch of enquiry using an inductive (discovery) method of learning. The methods chosen for doing so are:

Literature review: An adaption of *grounded theory* (Glaser and Strauss, 1967) to examine both the industrial and academic literature, extracting key information, codifying, classifying and collating information and combining this with unstructured memos to produce a comprehensive store of information about data security topics.

Questionnaire: The production, administration and analysis of a questionnaire aimed at a small population of data professionals with the aim of collecting their opinions on the imperatives in data security – this includes which techniques are worthwhile, current sources of information, and perceived gaps. This is a structured questionnaire using a combination of multiple-choice Likert scales and freeform answer collection.

Interviews: Multiple interviews with a range of data specialists aimed at a) discovering more about any conclusions reached from the questionnaire, b) triangulating and validating conclusions drawn from the literature review and c) discovering new information about the challenges in data security.

Figure 2 illustrates the principal tasks:

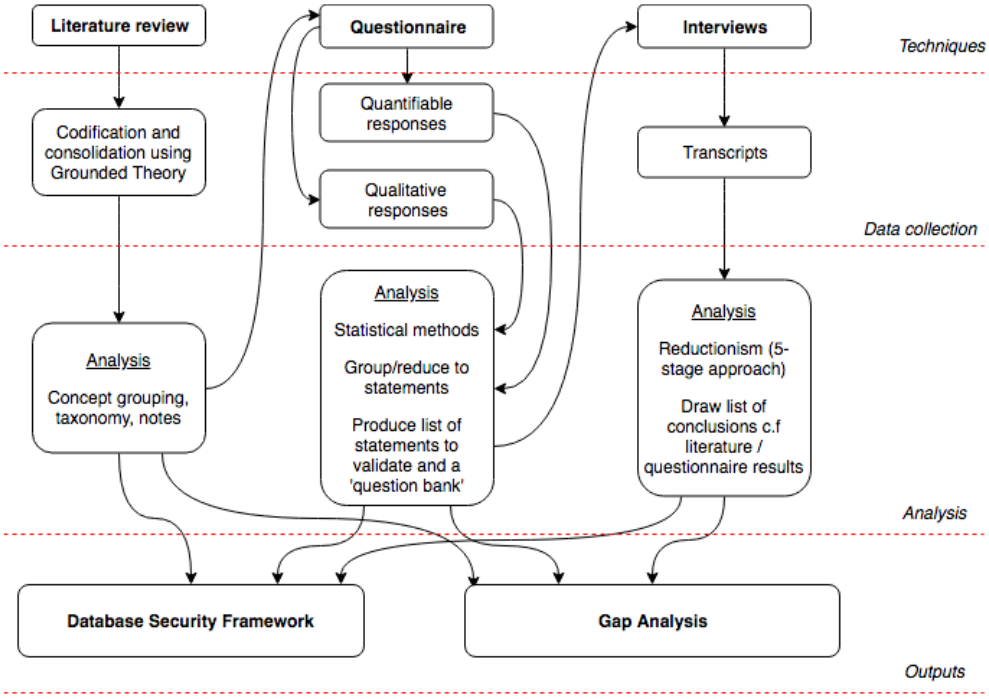


Figure 2: From techniques to deliverables

Data Collection and Analysis

For the *literature review*, more than 350 sources of information, a combination of academic literature, IT governance standards and software technical literature, were considered of which around 150 sources were deemed important and extracts and key themes recorded. For the *questionnaire*, a limited number of respondents meant analysis moved from a statistical foundation to acknowledging that sample scarcity meant a more anecdotal or narrative interpretation was required. *Interviews* were more successful, with 3 in-depth interviews yielding a large amount of information which was summarised, codified and analysed using a 5-step process for qualitative data analysis based on Boyatzis (1998).

The data analysis phase looked to group and analyse data from all three sources. In doing so, a series of important insights into data security were confirmed which form the general conclusions of this research, below. In addition, sufficient detail was extracted from all three methods of enquiry to form a taxonomy of topics which became the topic headings of the Data Security Framework. Alongside this content, the structure of the DSF was defined as a two-layer hierarchy with attribute labels defining the detail for each topic. The GAP analysis showed particular areas where industry and academia have diverged, and this is presented as a series of commentaries, broken down by topic. Finally, the impact of technological progress and cultural observations were considered.

Conclusions and Future Work

In this section, 3 implementations of the Data Security Framework are detailed – one in a relational database, another in a non-relational database and the third as an actual implementation as a Windows Help file, together with full topic record examples. The diagram below shows a screenshot of the DSF in action:

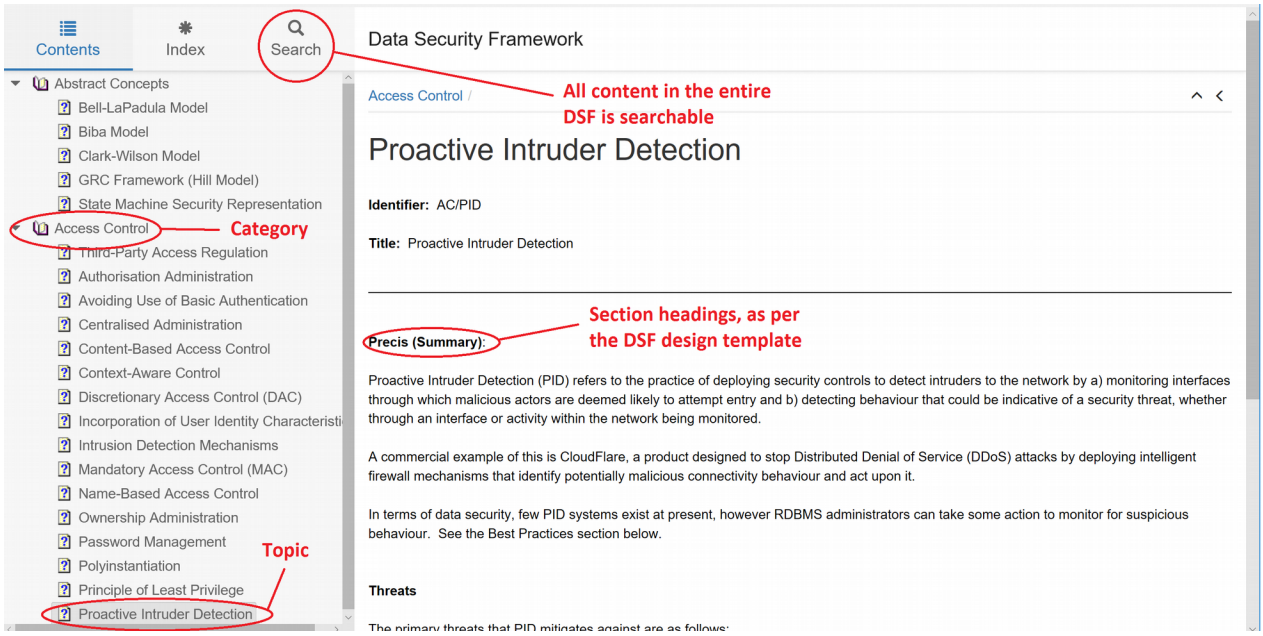


Figure 3: The DSF implemented as a Windows Help file

The GAP analysis is completed by aggregating the findings of the data collection phase to two SWOT analyses which provide illustrative examples of areas where industry can apply academic ideas and where academic researchers can take inspiration from industry examples. Finally, a set of general conclusions are drawn, a selection of which follow below:

- The persistent evidence of data breaches that affect organisations show that data security remains a current concern;
- Practitioners use similar best practices derived largely from shared knowledge and practitioner literature, but lack a unified security framework;
- Research material is broad and deep when considering long-standing information security concepts such as access control, but less so for newer technologies and in some areas, research is not keeping up with industry developments;
- The Data Security Framework produced provides a unified reference model for practitioners and researchers to determine the best course of action in applying security controls to their organisations or scoping out new areas for research, respectively;

- The pace of technological development is causing serious issues in ensuring good data security throughout the design and development lifecycle;
- More validation is required to make sure that the DSF is a viable and useful tool for users.

References

Bentley, C. (2012). *PRINCE2: A Practical Handbook*. London: Routledge.

Boyatzis, R.E. (1998). *Transforming qualitative information: Thematic analysis and code development*. London: Sage.

Olivier, M.S. (2002) 'Database privacy: balancing confidentiality, integrity and availability', ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp.20-27 [Online]. Available at: http://mo.co.za/open/dbpriv.pdf (Accessed 25 Oct 2017).

Solid IT Consulting and Software Development gmbh (2018). 'Knowledge Base of Relational and NoSQL Database Management Systems: DB Engines Ranking' [Online]. Available at: https://db-engines.com/en/ranking (Accessed 28 Aug 18).

Appendix A – Interview Question Bank

These are a set of 10 open-ended questions that result from the analysis of the questionnaire outputs, and from themes in the literature (secondary research). Questions are also based on the ontology of themes produced from the literature review. Each question has clarification questions indicated in *italics* which can be used to elicit more detailed responses, if necessary.

1) What techniques do you feel make good security principles in data management?
   For example, managing user permissions or auditing failed logins?
2) Which do you feel is the biggest risk in data security – internal threats, or external threats?
   a. *Why do you think that?*
   b. *Which is more likely to occur – internal or external threats?*
3) Is there a gap between best practices, as you understand them, in data security and what is actually applied to database management systems?
   a. *Can you give me any examples?*
   b. *What do you think can be done about this?*
   c. *What are the causes of the gaps between best practice and implementation?*
4) Does your organisation adhere to ISO 27001 standards?
   a. *Yes – How effective do you feel this standard is in mitigating information security risks?*
   b. *Yes – What can be done to improve the standard?*
   c. *No – Is there any particular reason your organisation doesn't use this standard?*
   d. *No – What standards do you use (if any) in your information security management system?*
   e. *Don't know – skip question*
5) Do you think enough is being done in application development to enforce security-by-design?
   a. *May need defining – summarise as security features built into the product e.g. hashed, salted password management; encryption; user access permissions audited etc.*
   b. *What improvements do you think can be made in how security is managed in the end product?*
6) Thinking about threat identification, how does your organisation identify and deal with emerging threats to the data platforms?
   a. *May need example – e.g. patching of security vulnerabilities, disabling accounts of 'leavers'…*
   b. *Ask e.g. if regular retrospective security reviews of user permissions are carried out*
7) How important do you feel data security is to your organisation?
   a. *Why do you think that?*
   b. *Do you feel enough is being done to secure the data at your organisation?*
   c. *What improvements could be made, in your view?*
8) Where do you get information on best practices in database security?
   a. *May need examples – e.g. product documentation, blog posts, internal procedures…*
9) What new features would you like to see in database security management?
   a. *If no ideas forthcoming offer suggestions such as integration with password management systems, inclusion of two-factor authentication and talk about usefulness of each.*
10) If there was a resource available to help you manage data security, like a central best-practice guide, is this something you would use in your day-to-day role?
    a. *Talk about the form which this guide could take, what factors would increase the use of this guide.*
    b. *Potential follow up – are you aware of any current research in database security?*

Appendix B – Classification and Analysis of Interview Responses

The following table shows how key statements from the interview data were extracted and analysed to produce a list of classifications or keywords. These keywords became the basis for grouping the data into themes and producing a series of conjectures – statements that capture the essence of the interview information, and which helped inform the content of the Data Security Framework (the thematic grouping is presented in Section 4.3.4).

Question	Interview 1	Interview 2	Interview 3	Code/Concepts
(1) Techniques that make good security principles	"Obviously patching is quite important, keeping everything up to date". "Making sure that the users have the right permissions for what they want to do". "Keeping sysadmin passwords … in Keepass [a password management tool] and in the safe". "The management need to be more aware of what we do … I think sometimes we're seen as a blocker* to the users doing what they want." "There's a lot more we could be doing … the [core CRM application] security is bad … everyone connects using the same application username."	"People are the biggest problem … you can have the most secure system in the world but it doesn't stop people from writing down their passwords." "When we develop, we think about security as we go, so things like … connection strings are centralised, nothing's plaintext …" "We use SSL [encryption] for all our traffic, that's pretty much standard now." "I follow quite a lot of infosec leaders, like Troy Hunt and Brian Krebs [Krebs on Security].	"…passwords need a lot of thought. Lots of people use the same password across different websites, if one gets hacked they all do." "Security needs to be thought about at the organisational level."	Patching Correct user permissions Sysadmin password control Management buy-in Application logins Security awareness Access control Encryption in flight Community sources Password rotation
(2) Internal threats vs. external threats	"They're both about the same." "The infrastructure team look after the external stuff … firewalls and that, and network separation." "Internal threats are harder but we audit a lot of stuff … and anyone who leaves has their account disabled." "The IT team disable things like USB and we have [web filtering product] which stops things like people accessing Dropbox."	"Internal threats are much more difficult to manage, most of the developers have prod[uction] access so if they get [annoyed] … there's not much we could do to stop them from taking whatever they wanted." We use CloudFlare [a popular DDoS prevention service] to stop most external attacks and ACLs on the firewalls, the networks are separate too so we're quite well protected." "There's always going	"Internal threats are the most damaging, but using the right permissions on the organisation's resources can help limit it." "The [hacking community] are getting cleverer, bot attacks and things like that are all automated now. This notion of someone sat in a basement trying to get into your server doesn't exist any more." "Hackers have come a long way since War Games! [1980s film drama featuring a young computer	Internal threats difficult to counter Disable unused accounts Disable uncontrolled media Apply web filtering Procedures for managing employee access after leaving employment Traffic filtering Existential threat Sophistication

		to be hackers out there, they're older than the Internet."	hacker]"	
(3) Gap between best practices and implementation	"Yes, there's a lot we need to do. I want to put in TDE [Transparent Data Encryption, a method for encrypting data at rest in Microsoft SQL Server] but we're stopped from doing that at the moment." "Everyone's got their own version of best practices." "A lot of what's on the internet is wrong. Take [redacted], everything he writes just isn't accurate." "I try and make sure my team is pretty consistent … we don't have anything formal but we all know what's acceptable and what isn't, if we get any crazy requests they will either escalate to me or bounce it back."	"What best practices? [laughs]" "We have some development standards and everyone tries to stick to these, but there is a lot of deviation, especially from developers who have been here a long time." "Our data team seems to be busy all the time." "We've got a backlog on our [Kanban] board and a lot of that is technical debt, so we're aware of the issues."	No significant answers were given.	Encryption at rest Variation in best practices Accuracy of best practices Consistent approach within business Technical debt management
(4) ISO 27001	"Yes we do." "It doesn't really deal with database stuff, but things like password policies are common sense." "It's quite broad-brush, you'd be better off speaking to [redacted: the infrastructure team lead]." "I've never seen an auditor."	"I think so. I don't really get involved in that sort of thing." "We've got pentesters [penetration testers / ethical hackers] in regularly enough so we're on top of any problems, I think."	"Yes, the University has that one … we have to, there's a lot of personal data floating around." "Some of the faculty had an input in getting it sorted out."	Password policies Pen testing Compliance Knowledge sharing
(5) Security in application development	"I don't really know how to answer that!" "We don't store passwords in plain text or anything if that's what you're asking." "I think there's a lot we could be doing better. We need to overhaul [CRM system] so we have a better idea of who's doing what."	"Yes, we've got quite a lot of that. Like I said earlier about storing connection strings. And all our passwords are hashed and salted." "We use Agile so it's a bit fast and furious sometimes but we do have templates and the dev[elopment] standards." "A lot of users have access in our reporting layer, I've been involved in discussions about that, I think it needs an overhaul really."	"We don't do application development so that's a hard one to answer." "Security by design was a big thing a while ago … a lot of research focus. I know we teach it in our undergraduate programmes." "I know cloud services use API key pairs so it makes it harder for anyone to access platform services."	Encrypted passwords Accountability Local standards Relaxation of permissions Security by design Cloud security

(6) Emerging threats	"We're normally patched up to N-1 [a shorthand term meaning the last available patch before the latest version]." "I keep an eye on [popular IT news forum]. So for example we found out about the SP1 index bug* before it actually affected us and I was able to patch it out." "There isn't really much that changes in databases, we keep everything patched and the database systems are airgapped** with the outside world."	"All the big companies do a lot of work with bug bounties. Google have a really generous program. And Microsoft have Patch Tuesday." "We keep everything patched and upgraded as much as we can and there's a couple of infosec guys who are on top of things, presumably!"	"There's always new research, new angles on how software has been compromised." "Data's got a lot of value. The dark net is full of people trading stolen credit cards and things like that." "I don't know how our IT department manage emerging threats. I imagine it's all done with patching and upgrades. My laptop is always applying upgrades [laughs] normally during lectures."	Patching Keeping up-to-date with industry developments Separation of database servers from wider internet Actively searching for security problems. Always new research Growing value of data
(7) Importance of data security	"Pretty important, but it's not at the top of my list … keeping the lights on is my priority." "I think what's more important is making sure everything's available all the time … if [CRM product] goes down then I've got 300 people coming to me to complain." "Data security comes down to me really. Although there's general guidelines, like we have a clean desk policy and the printers are secured." "I don't think another policy would do much good. I just want people to have a better awareness of security really. We can't have a situation where everyone in the dev team knows the sa* password." "Training might work, but I don't know if there will be a lot of interest."	"Of course it's important. How many companies have had a data breach? Someone told me that those companies who claim they haven't have, but just haven't found out about it yet!" "We've got dedicated teams for data and infrastructure so between them there's plenty of visibility." "Identity theft is a problem. It happened to me once … [details redacted] … took ages to get the money back and there was no explanation what happened either."	"It's of paramount importance. If security is compromised it could damage the whole business." "I don't like to think about what would happen if our student details were leaked online." "There's new encryption techniques coming out all the time. The problem is that computing power is increasing so rapidly that older algorithms, the ones that rely on short keys, can be brute-forced now, so there's a big push to move to longer keys."	Availability General, not specific, guidelines Security awareness User training Data breaches Dedicated engineering resource Business damage Older techniques are insecure
(9) New features	"Nothing really, I can't think of anything." "[Two-factor authentication] is a good idea but how would it work? … You can't have people getting a PIN code to access the database."	"That's the problem, I think there's too many new features now. There's so many different languages and tools out there no-one has a handle on what everything does." "It would be nice to be	"I'm not sure about that. I know relational databases have been around a long time. NoSQL databases might be a way forward."	Implementation challenges Hiding connection strings Two-factor authentication Password obfuscation

	"I wouldn't mind seeing something that could hide [SQL Server] passwords like Windows ones are hidden … you know what I mean? Integrated authentication or whatever it's called."*	able to connect to databases using class methods, instead of having to configure the SQL connection each time." "I heard something about SQL [databases] supporting Python now, is that true?"		NoSQL to replace structured databases Application support in RDBMSs Hard to assess/manage introduction of new technology
(10) Best practice guide for day-to-day use	"It doesn't sound like a bad idea. Is this something you're working on?" "I don't read any academic stuff, no. Only what's on the blog sites and things. Like [redacted]." "There's quite a lot of information already on Books Online."*	"We've got a best practice guide already, I'm not sure who would read another one?" "It could be handy for the DBAs"	"I think it's a good idea. There's no good central resource for … that. There's still lots of current research in journals like ToDS [Transactions on Database Systems]. The focus seems to be on AI these days." "It would have to be something a bit more than a technical manual. And there's so much research I'm not sure how feasible it would be."	Academic vs. industrial disconnect Information already exists Applicability No current central resource Lack of focus on security Applicability

Appendix C –Conjectures from Thematic Analysis of Interview Data

The following list of conjectures is the result of thematic analysis of the outcomes of the interview data as described in Section 4.2.2.

Theme / Concept	**Insight**
Knowledge management	A consistent approach can strengthen the effectiveness of good password management
Knowledge management	Better security awareness can aid adherence to password policies
Knowledge management	Having vague guidelines can cause ambiguity in implementation
Knowledge management	Having vague guidelines causes variation in best practices within an organisation, creating vulnerabilities
Knowledge management	Setting the correct user permissions is associated with good user access control
Knowledge management	Community sources can be valuable sources of information for data security matters
Knowledge management	Sharing knowledge within an organisation helps strengthen data security controls
Password management	Privileged credentials should be protected and policies created to avoid misuse
Password management	Passwords should be held in an encrypted form wherever possible
Password management	Password policies should exist which specify protections including regular rotation and obfuscation
Password management	Connection strings should be held securely and access controlled to ensure security
Barriers to understanding	To ensure accountability, put in place effective access controls. Likewise, a drive for accountability promotes good access control mechanisms
Barriers to understanding	Barriers to understanding lower the effectiveness of proactive measures to safeguard data security
Barriers to understanding	There is evidence of a disconnect between the academic research sphere and applicability of controls within industry
Barriers to understanding	Threats to an organisation are sophisticated and continually evolve, necessitating regular security reviews
Barriers to understanding	Internal threats are more difficult to counter than external threats due to the increased trust of an internal malicious actor

Influences	Data is perceived as continually growing in value meaning the likelihood of data security incidents increases
Influences	Older security control techniques (for example, encryption with short keys) is increasingly inadequate to counter evolving threats
Influences	Cloud security is an issue for organisations wishing to migrate to cloud computing but which comes with an expanded and evolving attack surface
Influences	NoSQL is widely seen as a potential replacement for relational databases due to the advantages given by different paradigms
Proactive measures	Many practical security controls exist including traffic filtering, connection string obfuscation, encryption and separation of duties
Proactive measures	Security by Design is an important concept that can reduce the likelihood that an attack will be successful by considering security throughout the design cycle
Proactive measures	Encryption is a fundamental cornerstone of good data security
Proactive measures	The provision of dedicated engineering resource is a necessity to help ensure good data security
Access control	Relaxing permissions for ease of use can have a detrimental effect by increasing the risk of unauthorised data access
Access control	Application logins should be managed to ensure confidentiality

Appendix D – Example Data Security Framework Topic Record

This appendix is an expansion of the structure of the DSF as given in Section 4.3.2, with embedded content examples.

When considering the DSF the following definitions are given:

- Category: A grouping of one or more concepts into a single common theme
  - For example, 'Development' incorporates three concepts
- Concept: A single topic, idea or theory in the domain of database security belonging to a category
  - For example, 'Enforcing security-by-design' belongs to 'Development'

A category has the following attributes:

- ID: A unique numerical one-part identifier in the form *x*, where *x* is a positive integer greater than 0.
- Title: The title of the category
- Precis: A short summary of the category, typically a few sentences.

A concept has the following attributes. **Bold** text indicates a linked concept.

- ID: A unique numerical two-part identifier in the form *x.y*, where *x* is the parent category identifier and *y* is a positive integer greater than 0.
  - For example, 2.9 or 3.13.
- Title: The title of the concept
- Precis: A short summary of the concept, typically a few sentences, relevant to database security.
  - For example, "'Security-by-design' means to keep security as a core goal during the development or administration of any application or process, rather than as an addition once finished. In this way, the impacts of threats and the attack surface are minimised. This is achieved using techniques during development that promote security, such as secure communication protocols, encryption, and defensive programming (e.g. for edge case handling), or technologies that promote security such as **contained databases**."
- Threats: A description of the attack surface that relates to the concept (if any); alternatively, how the threat profile is affected.
  - For example, "'Enforcing security-by-design' is to counter the threats of unauthorised access through leveraging unchecked vulnerabilities in an application caused by the omission of security measures during design. An example of this is the threat of **SQL injection** through lack of input validation and the use of dynamic SQL".
- Best practices: A brief description of best practices that relate to the concept.
  - For example, 'Enforcing security-by-design': "Security-by-design (SBD) can be instilled through communication with developers by team leaders; workshops around defensive programming; the enforcement of the use of secure technologies (such as SSL) and **encryption** within intra-team coding standards; the education and training of developers in associated technologies; and the regular audit, retrospective, peer-review and/or penetrative testing of systems that are affected. Example: Database schemas can be held in source control systems as part of a CI/CD release management system. To comply with security-by-design, one method would be to store

user permissions in source control which can then be dropped/recreated per release, thus being auditable and current."

- Academic summary: A brief description of any relevant academic research in this area.

For example, 'Enforcing security-by-design': "Dougherty et al. (2009) describe a set of general SBD design solutions that can be used in a variety of situations. In OWASP (2018), ten separate security principles govern SBD and many are applicable to DB security, including the **principle of least privilege** and minimise attack surface area. Vergenadis et al. (2017) present 'PaaSword', a framework for regulating access control to cloud resources (including data) using an alternative to **MAC** and **DAC**, 'Attribute-Based Access Control (ABAC)'.

Appendix E – SQL Code Example for DSF Implementation

This appendix contains an example SQL implementation of the schema presented in Section 5.1.3.

```
CREATE TABLE dbo.Category (
        CategoryID INT PRIMARY KEY NOT NULL,
        CategoryName VARCHAR(255) NOT NULL )

CREATE TABLE dbo.Topic (
        TopicIdentifier VARCHAR(10) PRIMARY KEY NOT NULL,
        TopicName VARCHAR(255) NOT NULL,
        Precis VARCHAR(MAX) NOT NULL,
        BestPractices VARCHAR(MAX) NOT NULL,
        AcademicSummary VARCHAR(MAX) NOT NULL )

CREATE TABLE dbo.CategoryToTopic (
        LinkID INT NOT NULL PRIMARY KEY,
        CategoryID INT NOT NULL FOREIGN KEY REFERENCES dbo.Category (CategoryID),
        TopicID VARCHAR(10) NOT NULL FOREIGN KEY REFERENCES dbo.Topic (TopicIdentifier) )

CREATE TABLE AcademicReferences (
        ReferenceID INT PRIMARY KEY NOT NULL,
        ReferenceText VARCHAR(255) NOT NULL )

CREATE TABLE dbo.Threat (
        ThreatID INT PRIMARY KEY NOT NULL,
        ThreatDescription VARCHAR(MAX) NOT NULL )

CREATE TABLE dbo.TopicsToThreats (
        LinkID INT PRIMARY KEY NOT NULL,
        TopicIdentifier VARCHAR(10) NOT NULL FOREIGN KEY REFERENCES dbo.Topic (TopicIdentifier),
        ThreatID INT NOT NULL FOREIGN KEY REFERENCES dbo.Threat (ThreatID) )

CREATE TABLE dbo.RelatedTopics (
        LinkID INT PRIMARY KEY NOT NULL,
        TopicFrom VARCHAR(10) NOT NULL FOREIGN KEY REFERENCES dbo.Topic (TopicIdentifier),
        TopicTo VARCHAR(10) NOT NULL FOREIGN KEY REFERENCES dbo.Topic (TopicIdentifier) )

CREATE TABLE dbo.TopicsToReferences (
        LinkID INT NOT NULL PRIMARY KEY,
        TopicIdentifier VARCHAR(10) NOT NULL FOREIGN KEY REFERENCES dbo.Topic (TopicIdentifier),
        ReferenceID INT NOT NULL FOREIGN KEY REFERENCES dbo.AcademicReferences (ReferenceID)
        )
```

Appendix F – GAP Analysis Topic Rankings

This is a continuation of the GAP analysis presented in Section 4.3.3.

The chart below ranks each topic as Good (green), Average (amber) or Poor (red) where grey is N/A.

The relevant questions that produce the ranking are, for each domain:

Industrial:
In the industrial sphere, how well is the topic understood and implemented, as it applies to data security?

Academic:

In the academic sphere, how extensively is the topic covered in the literature?

Concept	UQID	Topic	Industrial	Academic
Abstract Concepts	1.1	Bell-LaPadula Model		
	1.2	Biba Model		
	1.3	Clark-Wilson Model		
	1.4	GRC Framework (Hill Model)		
	1.5	State machine security representation		
Access Control	2.1	3rd-party access regulation		
	2.2	Authorisation administration		
	2.3	Avoiding use of basic authentication		
	2.4	Centralised administration		
	2.5	Content-based access control		
	2.6	Context-aware control		
	2.7	Discretionary access control		
	2.8	Incorporation of user identity characteristics		
	2.9	Intrusion detection mechanisms		
	2.10	MAC / LaPadula - no hierarchical traversal		
	2.11	Mandatory access control		
	2.12	Name-based access control		
	2.13	Ownership administration		
	2.14	Password management		
	2.15	Password policy		
	2.16	Polyinstantiation		
	2.17	Principle of least privilege		
	2.18	Proactive intruder detection		
	2.19	Restriction of sysadmin accounts		
	2.20	Retrospective review		
	2.21	Separation of duties		
	2.22	Temporal factors		
Auditing	3.1	Audit changes to data		
	3.2	Audit changes to schemas		
	3.3	Audit failed logins		
Availability	4.1	Cloud databases (DaaS)		
	4.2	Database unavailability		
	4.3	DoS / DDoS protection		
Business View	5.1	Agile development methodology		
	5.2	Disaster recovery planning		
	5.3	Formal user authorisation procedures		
	5.4	Insurance		
	5.5	Legacy system security requirements		
	5.6	Proactive over reactive security		
	5.7	News dissemination through external sources		
	5.8	Operational recovery planning		
	5.9	Regularly review ISMS documentation		
	5.10	Risk management		
	5.11	Security incident management response		
	5.12	Standards not practicable		
	5.13	User training		
Confidentiality	6.1	Air-gapped systems		
	6.2	API/key-based access control		

	6.3	Data mining		
	6.4	Hippocratic (Agarwal) databases		
	6.5	Multi-level relations		
	6.6	Negative authorisation		
	6.7	Password rotation		
	6.8	Row-level filtering		
	6.9	View-based access		
	6.10	Virtual private database		
Data Categorisation	7.1	Data labelling		
Development	8.1	Developer training in data security		
	8.2	Enforcing security by design		
	8.3	Separation of production / test data		
Encryption	9.1	Certificate and key management		
	9.2	Common Criteria Certification		
	9.3	Cryptography types		
	9.4	Encryption-at-rest		
	9.5	Encryption-in-flight		
	9.6	FIPS 140-2 standard		
	9.7	Use of standards		
Environmental	10.1	3rd-party security requirements		
	10.2	BYOD policies		
	10.3	Data-sharing policies for staff		
	10.4	Misinformation		
	10.5	Patching and updates		
	10.6	Physical security		
	10.7	Pre-existing documentation		
Exploitation	11.1	CSV injection		
	11.2	Homoglyphic attacks		
	11.3	Output leakage		
	11.4	SQL injection		
Integrity	12.1	Behavioural analysis		
	12.2	Anti-corruption measures		
	12.3	Data completeness		
	12.4	Data quality		
	12.5	Digital signatures		
	12.6	Logging		
	12.7	Loss of trust in data		
	12.8	Security of backups		
	12.9	Semantic integrity		
New Techniques	13.1	Automated disconnection		
	13.2	Automated vulnerability reporting		
	13.3	Biometric authentication		
	13.4	Blockchain integration		
	13.5	CAPTCHA-style authentication		
	13.6	Cross-platform security integration		
	13.7	SSO integration		
	13.8	Password management tool integration		
	13.9	AI / machine-learning led security analysis		
	13.10	Separation of security from RDBMS		
	13.11	Two-factor authentication		
Obfuscation	14.1	Connection string obfuscation		
	14.2	Data masking		
	14.3	Negative databases		
	14.4	Privacy concerns		
Role-based Authentication	15.1	Application vs. user roles		
	15.2	Automated role management		
	15.3	Object-based authentication		
	15.4	Task-based authentication		
Standardisation / Compliance	16.1	COBIT		
	16.2	Internal policy compliance		
	16.3	GDPR		
	16.4	Insufficient control specifications		
	16.5	ISO 27001		
	16.6	NIST 800-53		
	16.7	PCI-DSS		
	16.8	Sarbanes-Oxley		

Appendix G – Full GAP Analysis Outcomes

The following table is a continuation of the GAP analysis presented in Section 5.1.4.

Industry

Category	*Topic*	*Description*	*Applicability*
Abstract Concepts	Bell-LaPadula Model	This model emphasises labelling specific data items and setting security access on an item-by-item basis.	Currently the opposite approach is used in RDBMSs, a discretionary access control model where access is granted to whole objects based on their owner – subject – and this access is transferable. The BP model is manifest to a small degree in row-level filtering but much more could be done to use more granular security in RDBMSs in this manner.
Abstract Concepts	GRC Framework (Hill Model)	An overall framework for Governance, Risk and Compliance that aims for horizontal integration of business strategy, process and risk management to further the goals of the organisation. Applicable to data security as any data security strategy under this model would need alignment with the rest of the business.	Data security governed by GRC would see policies and controls shaped by business requirements rather than arbitrarily chosen, which would add value by strengthening an organisation's data security perimeter. In literature GRC is covered well with several principal authors in the field.
Access Control	Avoiding use of basic authentication	Ensuring that authentication is never carried out in plaintext and is preferably carried out in conjunction with a domain controller to authenticate users using SSO / ISAM technology.	Authentication is covered extensively in the literature but RDBMS platforms still allow support, e.g. SQL authentication in SQL Server and e-mail server communication tooling, and standard username/password capabilities in Oracle and MySQL.
	Content-based access control	Similar to Bell-LaPadula, this means to restrict access to users based on the content of an object rather than the object itself. However, the labelling conventions in BP do not apply.	Not used in RDBMS systems but potential to apply to industry especially in cases where data with mixed access levels is combined in single objects.
	Incorporation of user identity characteristics	This is a superset incorporating biometrics, two-factor authentication and any technology that can triangulate a user's identity with a secondary method.	Current authentication mechanisms in RDBMSs are one-dimensional – key/pair, or username/password combinations without 2FA. Data platforms would benefit from a secondary factor.
	Intrusion detection mechanisms	Technology that can detect whether either an inbound request for connectivity, or an active connection to a data source, is valid and a verified user.	This can be implemented at present in the network/application layer but would be a useful feature at the data platform perimeter to stop internal threats.
	Polyinstantiation	A technique with good research pedigree but not used in RDBMS platforms where relations exist as classes and multiple instances of the class – copies of the table – can be instantiated. Currently RDBMSs work on a transactional, lock-based basis.	A new paradigm for relational platforms but used with some success in NoSQL, notably Mongo, as 'sharding'. Not compatible with ACID principles but application of these ideas could be valuable.
	Proactive intruder detection	An extension of intruder detection mechanisms, this is to act upon the discovery of an intruder by e.g. terminating the connection, preventing further connection attempts and alerting users.	Similar to intrusion detection mechanisms, limited application at present in RDBMS systems. Development of these ideas could lead to better data security.
	Retrospective review	A technique native to Agile meaning to look back over a sprint or release and determine what lessons can be learned.	It is uncommon to regularly review recent security logs or changes from a data security perspective, and better awareness of the benefits of doing so may be

			beneficial.
	Temporal Factors	Temporal features in RDBMS products are slowly coming into play with support in the latest versions of the major platforms. However temporal features are rarely used and not understood.	Temporal features are not often used especially for security and authentication, for example by adding expiry dates to user accounts. This is a powerful way of ensuring time-limited access where appropriate.
Business View	Regularly review ISMS documentation	Part of the formal requirement of ISO 27001, ISMS procedures should be reviewed and updated regularly, normally at least on an annual basis. This applies to data security procedures also.	Data security should be included fully in the ISMS and responsibility for regular review given to the database administrators.
	User training	Users should be trained on the importance of data security and what they can do to support this aim.	Users already receive training, generally, for general IT security but delivery of training on the value and protection of data could help mitigate risks for organisations.
Confidentiality	Hippocratic databases	Based on current database design patterns, this is a privacy-by-design database which incorporates augmentations like data cleansers and validators, prioritising the data over the data structure, and particularly strong audit capabilities.	These are a paradigm that have not been developed in industry. However with the advent of GDPR in Europe, there is more focus on data privacy and re-examining these ideas could be worthwhile.
	Negative authorisation	A method of incorporating counterfeit credential data alongside real credential data in production databases so that the impact of breach is lessened as the value of the data is diminished. A form of data obfuscation. Genuine applications would be able to filter out genuine data through additional authentication mechanisms i.e. key/pairs / 2FA.	Obfuscation of credentials already takes place (hashed, salted passwords for example) but obfuscation through data hiding may add value.
Data categorisation	Data labelling	A subset of Bell-LaPadula and related to content-based access control, the act of labelling data with metadata that describes it, in order to better regulate who has access to that data.	Not currently supported by the relational model but some metadata supported by individual platforms, i.e. Extended Properties in MSSQL at the object level. Can also be implemented by design where sensitive data is kept in different relations.
Development	Developer training in data security	Developers should be trained on the importance of data security and what they can do to support this aim. This is closely related to Security by Design.	Little evidence that application developers are given data security training.
Encryption	Common Criteria Certification	An international standard for computer security. Most RDBMSs have CC certification including Oracle and SQL Server. However the technical detail of the standard is largely absent from practitioner literature indicating an absence of awareness.	Better awareness of the benefits of CC could be beneficial to enhance data security knowledge across the industry. Lessons from the technical detail of CC may also benefit application development.
	Encryption-at-rest	Meaning that data that is persisted is encrypted, so that access to that data without the appropriate keys yields unreadable data with no value.	Supported in RDBMSs through e.g. Transparent Data Encryption (TDE). However not a given and not enforced, also only available with certain software editions.
	Use of standards	Although there are some standards available (ISO 27001 for example), it is indicated that in many cases these standards are not used to their fullest extent, or are absent. This is a particularly strong theme in the research outcomes.	Evidence that ISO 27001 is a framework rather than a control specification meaning practical utility to developers is limited.
Exploitation	CSV injection	A method to inject malicious code	Can be countered through data input

		into a database by tampering with raw data in CSV (comma-separated value) files to get the query execution process to execute arbitrary code.	validation techniques and avoiding the use of bad practices in SQL development, for example the use of dynamic SQL.
	Homoglyphic attacks	Similar to CSV injection, this is the replacement of characters or character strings with similar character strings with the intention to disrupt the database engine or inject arbitrary code.	As per CSV injection, data validation can help mitigate the impact of this exploitation technique.
	Output leakage	This is the concept that authentication mechanisms can leak information to unauthorised users via the error messages or behaviour that the authentication mechanism exhibits.	Ties into Security by Design. Output leakage can make attacks more likely since the actor can verify or eliminate various attack vectors based on the error output from the authentication process. Much has been written in the literature about this and there exists good guidance for writing opaque interfaces to counter this threat.
	Behavioural analysis	Meaning to examine and analyse the behaviour of a user or group to determine a common usage pattern and to raise alerts when that usage pattern is not adhered to. Notably in use in web application authentication, for example when a US-based Google user logs in from another country, Google will flag the login as potentially fraudulent and alert the real user via email, or may refuse the login altogether.	This is an area frequently explored in the literature especially with the rising popularity of machine learning solutions, but not generally employed in the field of data security.
	Data quality	RDBMS systems are concerned with the schemata of data rather than the content. Although this is changing with the introduction of data cleansing tools (Master Data Services for SQL Server, for example), emphasis continues to remain on data access rather than data usefulness. Data quality means how accurate, valuable and timely the data is to the organisation.	The importance of data quality is recognised through the 'Vs' of big data – particularly value and veracity. However the relational model in its pure form is unconcerned with the quality of data beyond its adherence to the type of the column in question. More could be done to build quality assurance into the relational model.
New Techniques	Automated disconnection	A method for disconnecting a connection automatically if undesirable behaviour is observed.	System timeouts exist but are based on connection TTL rather than any behavioural activity. There are opportunities to e.g. monitor for patterns of queries used by pen-testers or hackers running scripts to collect system information and close these connections automatically. While these would yield false positives, they could help prevent unauthorised access.
	Automated vulnerability reporting	A method for detecting potential vulnerabilities in software automatically by the self-analysis of crash reports, pen-tests, user behaviour and log traffic.	Some coverage in the literature and not currently deployed in industry. Vulnerabilities are found through bug-bounty programs and by the software suppliers, and patched manually.
	Biometric authentication	The authentication of a user by biological factors such as fingerprint or retinal scan.	No current mechanisms in place to do this for RDBMS systems and rare in the application space, but the technology is already fully developed and well-covered in the literature.
	Blockchain integration	Using a blockchain ledger, a series of co-dependent transactions based on prime number calculations, to	A new technology stemming from the cryptocurrency Bitcoin, blockchain enables fully-trusted verification of transactions.

		guarantee integrity of a series of transactions.	Plenty of literature coverage especially for transactional systems like RDBMSs. Industry also proceeding with implementations but no evidence of being on roadmap for RDBMSs.
	CAPTCHA-style authentication	A variant of two-factor authentication but one that can be done on a single device and requires human intelligence to solve.	See 2FA, CAPTCHA is an additional mechanism that could be used to verify users during authentication.
	Password management tool integration	The integration of password management tools (such as KeePass, and others) into data platforms so at no point are system passwords exposed to the user.	Tools such as KeePass are widely used and other major platforms incorporate key stores, such as Apple's Keychain. Little support exists (except for Oracle) for native key stores in the RDBMS model.
	AI / machine-learning led security analysis	The application of AI / ML techniques to security traffic to learn common access patterns and alert if extraordinary behaviour is observed.	This technology already exists and is applied to network traffic most frequently, with coverage in the literature, but is not generally applied to the RDBMS interface.
	Separation of security from RDBMS	RDBMS platforms currently include security as a core feature through the use of roles, logins, users, owners and credentials. Separating security puts the onus of authentication on the calling application rather than the RDBMS.	Some developments in industry including the idea of 'contained databases' (MSSQL) which, rather than separate security aspects from design aspects, fully incorporate them into a single container. Security models are well-covered in the literature.
	Two-factor authentication	To authenticate a user using at least two distinct methods which combine to triangulate a user's identity. Typically achieved using a secondary device such as a mobile phone.	RDBMS interfaces typically use single-factor authentication but the opportunity exists to use 2FA – currently deployed in the latest versions of email client Outlook, for example, and in most cloud services. This would shut down attack vectors such as the use of stolen passwords, for example.
Obfuscation	Negative databases	Related to negative authorisation, negative databases have counterfeit data alongside real data and real data is filtered out through the use of supplementary authentication mechanisms.	Extends the principle of negative authentication to the whole database. Some good coverage in the literature but does not appear to have been taken up at all in industry.
Role-based authentication	Automated role management	Role management in RDBMS systems is currently carried out by the database administrator. Automated role management would integrate with other data sources e.g. organisational charts to automatically assign roles to users based on their level of agreed access.	Would require tighter integration with a staff management software system to implement permissions based on external data. Not currently used but the idea has been explored in literature.
	Task-based authentication	A mechanism to assign only the required level of permissions per task for the task to be successful.	Would require foreknowledge of the task and a categorisation system to be aware of which permissions each task would require. Not suitable for bespoke systems. Would have the disadvantage of needing a controller that could arbitrarily assign permissions, but could make RDBMS connections safer. Some exploratory literature is available.
Standardisation / Compliance	COBIT	A business management framework concerned with IT governance standards.	Much literature on COBIT but not a framework that is commonly associated with RDBMS systems. Some scope here for using COBIT alongside ISO 27001, but may suffer from lack of details in controls.
	Insufficient control	Observed particularly with ISO	This research aims to show how specific

	specifications	27001, where control mechanisms are in place but not in enough specific technical detail to enable the technology teams to be able to implement them in a consistent manner.	controls can be documented for every variety of data security issue and used alongside governance standards. A standard set of controls for industry would be beneficial. There are many examples in the literature of specific deployments of security control groups applicable to certain sectors or systems only (e.g. healthcare applications) but no whole-industry attempt.

Academic

Category	*Topic*	*Description*	*Research Directions*
Access Control	Intrusion detection mechanisms	Technology that can detect whether either an inbound request for connectivity, or an active connection to a data source, is valid and a verified user.	This is an overarching topic for methods like behavioural analysis and there is some generic literature available, but little when considering interfaces to data systems. More research into the possibilities of detecting e.g. aberrant query patterns could be beneficial.
Availability	Cloud database(s)	Data security in cloud databases is a topic that benefits from extensive industry application and development but is applied in different ways depending on the platform. Typically reliant on application API/key pairs and traditional username/password credentials.	Cloud computing is a relatively new paradigm and research appears to lag behind latest developments. Although there exist standards like ISO 27017 which cover cloud-based systems, developments in industry continue apace, and there is the problem of divergence between platforms. More research into mitigating security issues would benefit the industry as a whole.
Business view	Standards not practicable	Where standards are defined at a detailed level but are not practicable in the context of the systems to which they should be applied. Applies particularly to legacy data systems such as early versions of Oracle, MSSQL etc. where controls like encryption are not supported.	This is a little-regarded problem in the literature but can exacerbate risk in the industry. Often legacy systems are critical to businesses, although keeping systems updated is a priority often older systems must be maintained and the security surrounding them becomes outdated. More research is required in how to maintain adequate security in these instances.
Confidentiality	Virtual private database	A form of data masking and supported very well in Oracle-based systems, VPDs hide data from a 'master' database and present it in a series of child databases depending on the definition of the objects to be displayed / hidden. For example a sales VPD might show only sales data from a larger database which also contains personnel data.	Implemented across industry although not ubiquitously. Achievable at the object level in other major platforms. The general idea has some research coverage but much more could be done in this area, for example by extending the principle down to the set-theoretic level, applying VPDs, which are essentially collections of views, to the individual relation. Other topologies are also possible.
Exploitation	CSV injection	A method to inject malicious code into a database by tampering with raw data in CSV (comma-separated value) files to get the query execution process to execute arbitrary code.	There is very little literature coverage of this method of input poisoning, although the related topic of SQL injection (normally through more direct interfaces) is well-covered. The industry is generally aware of the risk of CSV injection and can mitigate through data validation.

New Techniques	Automated disconnection	A method for disconnecting a connection automatically if undesirable behaviour is observed.	Linking to intrusion detection mechanisms, there is little research in how undesirable query behaviour can be discovered and responded to.
	Automated vulnerability reporting	A method for detecting potential vulnerabilities in software automatically by the self-analysis of crash reports, pen-tests, user behaviour and log traffic.	Machine learning is undergoing a revival and the way that big data is processed to look for patterns and abnormalities is progressing. However, there is little research into how this kind of insight mechanism can be applied to RDBMS platforms.
	Password management tool integration	The integration of password management tools (such as KeePass, and others) into data platforms so at no point are system passwords exposed to the user.	There is very little research into this area and how supplementary key stores can be integrated to RDBMS authentication mechanisms. This could be an exciting area of research.
	Task-based authentication	A mechanism to assign only the required level of permissions per task for the task to be successful.	This is a paradigm shift for relational authentication since permission sets are static at present. For permissions to alter in accordance with need would require a rethink of the principles of role-based authentication and could be promising.
Standardisation / Compliance	Insufficient control specifications	Observed particularly with ISO 27001, where control mechanisms are in place but not in enough specific technical detail to enable the technology teams to be able to implement them in a consistent manner.	There is some coverage of this topic in the literature, particularly from papers which seek to reinforce the justification for their control-based research into information security by examining the shortcomings of existing standards. However, there is a gap for an empirical review of standards effectiveness in organisations.