Big Data, Unstructured Data, and the Cloud: Perspectives on Internal Controls

David Simms

Haute Ecole de Commerce, University of Lausanne, Switzerland
david.simms@unil.ch

Abstract. The concepts of cloud computing and the use of Big Data have become two of the hottest topics in the world of corporate information systems in recent years. Organizations are always interested in initiatives that allow them to pay less attention to the more mundane areas of information system management, such as maintenance, capacity management, and storage management, and free up time and resources to concentrate on more strategic and tactical issues that are commonly perceived as being of higher value. Being able to mine and manipulate large and disparate datasets, without necessarily needing to pay excessive attention to the storage and management of all the data that are being used, sounds in theory like an ideal situation. A moment's consideration reveals, however, that the use of cloud computing services, like the use of outsourcing facilities, is not necessarily a panacea. Management will always retain responsibility for the confidentiality, integrity, and availability of its applications and data, and being able to develop the confidence that these issues have been addressed. Similarly, the use of Big Data approaches offers many advantages to the creative and the visionary, but such activities do require an appropriate understanding of risk and control issues.

1 Introduction

This chapter will set out the risks related to the management of data, with particular reference to the traditional security criteria of confidentiality, integrity, and availability, in the contexts of the wider use of unstructured data for the creation of value to the organization and of the use of Big Data to gain greater insights into the behaviors of markets, individuals, and organizations. Of particular interest are the questions of identifying what data are held internally that could be of value and of identifying external data sources, be these formal datasets or collections of data obtained from sources, such as social networks, for example, and how these disparate data collections can be linked and interrogated while ensuring data consistency and quality.

Much of the current debate around big data technologies and applications concerns the opportunities that these technologies can provide and where issues of security and management are addressed; there is a tendency for these to be considered somewhat in isolation. This chapter will set out the risks, both to the owners and the subjects of the data, of the use of these technologies from the perspectives of security, consistency, and compliance. It will illustrate the areas of concern, ranging from internal

F. Xhafa et al. (eds.), *Modeling and Processing for Next-Generation Big-Data Technologies*,
Modeling and Optimization in Science and Technologies 4, DOI: 10.1007/978-3-319-09177-8_13

requirements for proper management to external requirements on the part of regulators, governments, and industry bodies. The chapter will discuss the requirements that will need to be met in respect of internal control mechanisms and identify means by which compliance with these requirements can be demonstrated, both for internal management purposes and to satisfy the demands of third parties such as auditors and regulators.

The chapter will draw upon the author's wide experience of auditing the IT infrastructures of organizations of all sizes in describing the processes for identifying relevant risks and designing appropriate control mechanisms. The chapter will contain discussion of the standard frameworks for implementing and assessing controls over IT activities, of information processing and security requirements, objectives and criteria, and of monitoring, testing evaluating, and testing control activities. The author will also apply his insights into the strategies being followed, and initiatives being considered, by a range of organizations and corporations in order to illustrate how risks are changing as technologies change and how control activities need to develop in response. This analysis will be based upon the results of a survey performed within Switzerland in 2011 and 2012 that set out to establish an overview of how organizations viewed and understood both cloud technologies and the nature and value of the data that they held, and what impact these concepts and opportunities would have on their strategies, policies and procedures.

2 Preliminary Concepts and Definitions

Many of the terms used in this chapter are reasonably recent coinages and definitions can still be flexible and varied.

For the purposes of this chapter, we will follow Bernard Marr [1] with the definition that "Big data refers to our ability to collect and analyze the vast amounts of data we are now generating in the world. The ability to harness the ever-expanding amounts of data is completely transforming our ability to understand the world and everything within it." The fundamental idea is of the accumulation of datasets from different sources and of different types that can be exploited to yield insights.

According to an article by Mario Bojilov in the ISACA Now journal [2], the origins of the term come from a 2001 paper by Doug Laney of Meta Group. In the paper, Laney defines big data as datasets where the three Vs—volume, velocity, and variety—present specific challenges in managing these data sets.

Unstructured data as a concept has been identified and discussed since the 1990s but finding a definitive and all-encompassing definition of what this might be is surprisingly difficult. Unstructured data have been defined vaguely positively by Manyika et al [3] as "data that do not reside in fixed fields. Examples include free-form text (e.g., books, articles, body of e-mail messages), untagged audio, image and video data. Contrast with structured data and semi-structured data" which provides a definition but one which is rather more in respect of characteristics that the data do not possess rather than what they are.

To complicate matters slightly, Blumberg and Atre [4] discussed the basic problems inherent in the use of unstructured data a decade ago. They wrote "The term unstructured data can mean different things in different contexts" and that "a more accurate term for many of these data types might be semi-structured data because, with the exception of text documents, the formats of these documents generally conform to a standard that offers the option of meta data" (p. 42).

Cloud computing is described by NIST as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [5], while the Cloud, according to Manyika et al, is "a computing paradigm in which highly scalable computing resources, often configured as a distributed system, are provided as a service through a network." For corporate users, the immediate impact of using cloud services is not necessarily clearly distinguishable from that of using traditional outsourced services, with services and/or data being available through an external network connection. The key distinction is that unlike conventional outsourcing, where typically the service provider contracts to store, process, and manage data at a specific facility or group of facilities, in the context of the cloud the data could be stored anywhere, perhaps split into chunks, with no external visibility over how that was organized.

For the individual, private user, the cloud, represented by such services as Dropbox, iCloud, or Google Services, for example, is exactly as its name suggests, a virtual and distant and slightly opaque facility by which services are provided without significant identifying features or geographical links.

Internal Controls are the mechanisms – the policies, procedures, measures, and activities – employed by organizations to address the risks with which they are confronted. To define a little further, these activities fall into the framework of control objectives are the specific targets defined by management to evaluate the effectiveness of controls; a control objective for internal controls over a business activity or IT process will generally relate to a relevant and defined assertion and provide a criterion for evaluating whether the internal controls in place do actually provide reasonable assurance that an error or omission would be prevented or detected on a timely basis [6].

The traditional triad of information security objectives consists of Confidentiality, Integrity, and Availability [7].

- Confidentiality is the prevention of the unauthorized disclosure of information, providing the necessary level of security.
- Integrity is the prevention of the unauthorized modification of information, assuring the accuracy, and integrity of information.
- Availability is the prevention of the unauthorized withholding of information or resources, ensuring the reliable, and timely access to the information for authorized users.

To give a concrete example of how control objectives and internal controls fit together, an organization might be concerned about unauthorized access to its data. The control objective might be to ensure the confidentiality of the data, and one of many

possible control activities might be to perform regular reviews of the appropriateness of user access rights at the application level, to ensure that there are no redundant or unallocated accounts. Clearly in a well-controlled environment there will be a number of internal controls relating to each activity and each objective, and management will draw their comfort from the combined effectiveness of these controls.

At the same time there are a number of other objectives in respect of information security that may be of greater of lesser significance for different organizations, depending on the data they hold and process and the sectors in which they operate. These areas, which are included in standards such as those published by the ISO [8] and NIST [9], among others, include:

- Authenticity / authentication: having confidence as to the identity of users, senders, or receivers of information;
- Accountability: ensuring that the actions of an entity can be traced uniquely to that entity.
- Accuracy: having confidence in the contents of the data;
- Authority: the means of granting, maintaining and removing access rights to data;
- Non-repudiation: making it impossible for the person or entity who has initiated a transaction or a modification of data to deny responsibility after the event;
- Legality: knowing which measures are appropriate for the legal frameworks within which an organization is operating.

From a point of view of completeness, it should also be mentioned that there are a number of other classifications of information security criteria. In 2002, for example, Donn Parker proposed an extended model encompassing the classic CIA triad that he called the six atomic elements of information [10]. These elements are confidentiality, possession, integrity, authenticity, availability, and utility. Similarly, the OECD's Guidelines for the Security of Information Systems and Networks, first published in 1992 and revised in 2002 [11], set out nine generally accepted principles: Awareness, Responsibility, Response, Ethics, Democracy, Risk Assessment, Security Design and Implementation, Security Management, and Reassessment.

3 State of the Art

Cloud computing is a paradigm in computing that has emerged from the spread of high-speed networks, the steady increase in computing power, and the growth of the Internet. The interconnection of resources that may be separated by significant distances is allowing users access to what appears to be a single resource. As a result, there is an opportunity to outsource resource-intensive or resource-specific tasks to service providers who will deliver the service for a fee, often based on consumption, rather than developing and maintaining the infrastructure and competences in-house.

Neither of the central ideas in this model, distributed computing or outsourcing, is new. Distributing resources within and across networks has been performed since the days of mainframe computers, where data were entered on remote terminals and processed centrally, while outsourcing of computing activities and services has taken place for many years for strategic, operational, and financial reasons.

In technical literature reference is often made to cloud computing, grid computing, and utility computing; there is no real consensus over whether there is a distinction to be made between the three and, if so, whether the distinctions are clear or are rather subtle. In general, though, the models are identified by features, such as ubiquitous access, reliability, scalability, virtualization, the exchangeability of and independence from location considerations, and cost-effectiveness [12].

A key area of growth is that of the types of service that can be offered by providers. Historically providers typically offered remote data processing, storage, and systems management as major services, but the newer paradigm is to group offerings together as types of services: Software as a Service (SaaS), Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Service-Oriented Architecture (SOA). The underlying principle is that every facet of computing activities can be offered as services to consumers to be paid for on a usage basis. Computing is offered as a utility with billing based on consumption, with a corresponding shift toward IT spending being classified as expenditure on a service rather than infrastructure investment [13].

This paradigm has been likened to the electrical power grid, both in the impact it has on social and industrial development and because of its nature: elements of the grid can be owned and operated by different entities in different locations and might not share any physical characteristics, but the users, those who pay for the service, will only see a single interface or point of contact and will consume a consistent and homogenous service [14].

3.1 The Advantages of Cloud Services

The advantages of such services for certain users are, at least in principle, clear. The users need not worry about maintaining the application infrastructure, with all the operating system, database, and application patches and upgrades that are necessary. Nor need the users worry about data management practices such as backups or transfer between machines and environments. Both individuals using office-style applications and remote storage services and large corporations using large-scale applications remotely will see the benefits of avoiding questions of ongoing software compatibility and upgrade paths. Essentially these services are seen as an opportunity to avoid having to deal with certain aspects of the complexity involved in managing a technological infrastructure.

For corporate users there are also the advantages related to the use of scalable architectures. Instead of making potentially significant capital investments in hardware and software that might never be fully exploited and thus represent an unnecessary cost to the business, management will be tempted to subscribe to the model of being able to request and exploit additional resources when required, and for as long as required, on something like a Pay-As-You-Go basis. This has the potential to allow a much more accurate and dynamic management of costs while still permitting absolute flexibility in the access to and use of resources.

The advantages for service providers of operating in this sector also seem to be well established. Once a data center has been established and the reasonably fixed costs of operation have been established and incorporated into the business model, the

variable costs of service provision and marginal costs related to the acquisition of additional clients or providing additional services or additional capacity to existing clients should be straightforward to manage and reasonably simple to recover (and exceed, of course) through appropriate pricing. Thus the provision of such services can be viewed as a potentially lucrative business, with important initial investment but steady and permanent future revenue streams.

Conceptually, the techniques of cloud computing are not significantly different to those of traditional outsourcing of IT services. Many service providers offer outsourced data processing, system management, monitoring, and control services and these services are very popular with many organizations who either do not wish to have internal IT services and competences for organizational reasons or prefer to outsource for financial or logistical purposes.

Key elements of any outsourcing agreement are the contract terms and the service level agreements. With these terms, it is clear to both parties which services are being provided, what resources are being made available, how these resources are being managed, and how the quality of service can be evaluated and managed.

Many structured cloud services will provide such terms and conditions but might not be able to specify every element of interest to the customer, such as the precise location where data are stored or processed.

It is in the nature of any kind of outsourcing or service provision activity that both parties will wish to maximize their revenues and benefits from the agreement while at the same time accepting the minimum level of responsibility for addressing the risks involved and for handling any issues that arise. In the context of cloud computing, customers need to pay particular attention to the clear definition of roles and responsibilities in order to try to avoid situations of blame-shifting and cost avoiding should problems subsequently arise.

A key driver for the use of cloud facilities is costs: organizations might not wish to tie up capital in IT infrastructure that might not be used to capacity and which might only have a short life before obsolescence, when there might exist the possibility of renting services as a regular P&L charge. Along with the resource and competency questions, which of course also incur ongoing costs and can be expensive to update or replace, this has long been a prime mover for outsourcing services. Experience in the domain of outsourcing, however, reveals that the cost savings may not always be as significant as hoped for. As mentioned above, the prudent management team will look to ensure that its systems and data are secure and reliable by implementing additional internal controls to generate evidence that there are no weaknesses in the service provider's controls that can be or are being exploited. Typically, these internal controls will take the form of data and transaction analysis to identify exceptions, or the appointment of specialist staff that can manage the relationship with the service provider and ensure that trends are identified, that service levels are maintained, and that issues are identified, reported, and rectified. Very often the costs associated with implementing and operating such additional internal controls in an effective and robust manner are such that they can eat into the margins created by the whole outsourcing initiative.

3.2 The Disadvantages of Cloud Services

If the advantages for users of cloud services, as set out above, appear to be obvious, then the disadvantages are equally clear, particularly if viewed from a management and control perspective. The management of organizations always retains responsibility for the security and availability of their systems and data, in particular for the three key attributes of confidentiality, integrity, and availability. Ensuring that these attributes are understood, managed, and evaluated has traditionally been a significant challenge in systems management, even when systems and data are kept within a well defined and efficiently policed perimeter. Once this perimeter is extended outside the sphere of direct control of relevant management, the challenge becomes increasingly difficult.

In an article published in the Gartner blogs [15], Thomas Bittman argues that the widely used analogy of cloud computing to provision of electricity or water supplies is not particularly illustrative, for two reasons: "(1) Computing is a rapidly evolving technology, and (2) Service requirements vary widely for computing. Electricity production and distribution hasn't evolved much since the invention of AC that made distance distribution possible. How many forms of water are needed around the world? It's H2O – maybe it can be potable, purified, or come at a special temperature, but it's still pretty basic stuff."

His analogy is that of transmitted music: radio broadcasts of music began in 1916 and phonograph cylinders were used for storage, with the idea being that people would not wish to bear the expense of storing their own copies of music when it was available over the airwaves. But technologies have advanced, both for the broadcast and transmission of music and for the local storage and consumption. Individuals continue to be prepared to pay a premium, for infrastructure, material and content, for a quality and rapidity of service.

The choice between broadcast and local access to music is the same as the cloud computing question: it will depend on a number of factors and requirements that are constantly evolving and ultimately become a question of the best balance between costs, risks, and quality. A key element in the cloud computing choice will be assessing the development of requirements and adopting a strategy that will maximize the Return on Investment, however that is calculated.

Such analogies do reach their limits, however, because they do not consider the specific nature of the service provided. When plugging in a laptop in a foreign hotel room, the consumer broadly speaking does not care who has provided the electricity and the identity of the provider has no impact on the use of the service. Using cloud services is fundamentally different, though, for as soon as a provider is storing data or performing processing for a client, there emerge questions of the confidentiality and integrity of the systems and data as well as the availability of services.

A further aspect in which the analogy breaks down is in the area of the difficulty of changing approach once decisions have been taken. There is no question of not continuing to use electricity to power devices, but there will always be questions about the most efficient way to have access to and use IT infrastructure and resources. Moving from one cloud solution to another is likely to prove as complicated, if not more, than moving from one traditional outsourcing provider to another.

3.3 Data Storage and the Cloud

The opportunity to exploit the storage potential of the cloud can feasibly be viewed as an encouragement to organizations to place less priority on good practices in relation to data management. As storage space increases, so does the tendency simply to store all data rather than to classify, prioritize, archive, and delete them as appropriate, and there is nothing in the principles of cloud computing to reverse this tendency.

It is therefore reasonable to envisage situations in which individuals and organizations simply do not know what data they have placed in the cloud. There may be multiple copies of the same documents and datasets. There may be inconsistent and incoherent datasets. There may be quantities of unstructured data – data extracted from central systems and databases that are not in standard, easily recognized, and easily classified structures – that by their nature have not been classified and evaluated.

Of course, the question of unstructured and unmanaged data is not unique to the cloud: it exists in virtually all IT environments. Where it becomes particularly significant in the cloud environment, however, is as a consequence of the lack of visibility the data owners have over their data. Access to in-house, and properly secured outsourced data collections, can be defined, logged, and reviewed, at least in theory given sufficient resources and sufficient motivation on the part of the data owners. Once the data are in the cloud, however, the owners have very little visibility and control over the security of their data [16]. If this weakness is exacerbated by an absence of real knowledge of what data are out there, the risk of data loss and disclosure is increased.

The classical model for many businesses and other large organizations is to have one or several centralized key systems in which all of the organization's financial and operational activities are recorded. Typically, the data used in key applications and upon which users depend are stored in structured, centralized and at least theoretically well-controlled databases.

There is also, however, widespread use of noncentralized data repositories. Individual departments or users may have their own specific applications that do not fall into the overall organization-wide systems landscape and thus are subject to different standards and procedures for management and control. This is not to say that these systems and data are necessarily badly controlled, simply that management cannot necessarily be certain that standard policies and procedures are being applied.

3.4 The Use and Frequency of Unstructured Data

In the modern business environment great use is made of unstructured data in a variety of contexts. Typically users will extract data from central databases, often those underlying ERP systems, and then manipulate these data in a variety of ways as part of their business activities. Such data are thus often found in spreadsheets, user-built databases, and desktop or departmental server-based applications. These data can also be found in text files, pdfs, and even multimedia formats, depending on the use to which they have been put.

From an internal controls perspective the presence and use of unstructured data can pose numerous problems in respect of the confidentiality and integrity of data, two-thirds of the famous "CIA" triad of information security objectives (the third being availability). When data are located in a structured central database, they can, at least in theory, be controlled, managed, verified, and secured. Once extracted from the database, though, they can easily escape the internal control environment [17]. They can be used for decision taking without the assurance that they are still current, complete, or valid. They can also, depending on the security measures in place and the efficiency with which these are enforced, leave the organization easily, typically on the USB media that have become ubiquitous, on laptop hard drives, or even as attachments to emails.

4 Problems, Issues, and Challenges

There are several underlying problems related to the management of multiple datasets and of collections of unstructured data.

The fundamental problems related to the management and control of data stored in the Cloud are not dissimilar to those encountered when using outsourcing facilities or third-party service providers. Simply stated, as soon as data or applications are moved outside the perimeter of integrated and monitored internal controls, management have less control over their data.

Typically in an outsourced environment there are a number of ways for the organization purchasing the services to acquire at least a reasonable level of control and assurance. These can be described as obtaining comfort from audit procedures performed by the organization itself or by an audit entity specifically mandated to audit on its behalf; by obtaining a service auditor's report for the environment in question; or by implementing, performing, and monitoring the success of a series of internal control activities.

4.1 Self-performed Audit Procedures

The first of these methods is the most direct and is identical in form and process to the use of internal audit functions to evaluate and report upon business operations that are carried out in-house, as well as the use of the findings of external audits, where internal controls are often evaluated in the context of a controls-based audit. The audit team works with the organization's management to define scope and timeframe, documents the business procedures, identifies the key control activities, assesses these for both **design effectiveness** (whether as designed the control activities do address the risks to which they relate and the control objectives to which they correspond), and **operating effectiveness** (whether the control activities are actually working as designed, with necessary inputs being received, the control being performed, and appropriate conclusions being drawn and actions undertaken). Based on the results of these audit activities, management will be in a position to draw informed conclusions on whether or not internal controls are working as intended.

The right to perform such audits will need to be specified in the contract for service provision.

The difficulties of performing such reviews in an outsourced environment are many. First, the organization may not have an appropriately skilled, experienced, or available audit department – it is frequently the case that internal audit functions tend to have greater expertise in, and requirements to focus on, financial management issues, such as the use and disposal of assets, Value For Money, and purchase and inventory management. Detailed technical skills in the field of information systems are more likely to be found in larger and more IT-dependent organizations.

Secondly, it might be prohibitively expensive to mandate a third party to perform the necessary audit on their behalf.

Thirdly, it might be logistically difficult to identify all the areas in which processing or data storage or IT management tasks are performed at the service provider, especially if multiple sites in multiple locations are used. This will add to the complexity of scoping and scheduling such an audit.

Fourthly, but far from being the least significant issue, there is an impact of being audited on the service provider. In order to respond to audit questions, staff, and documentation need to be made available, and then resources need to be provided to assist with the testing of individual controls, extracting system information and reports, and explaining their contents. This can have multiple impacts on the service provider, taking up human and technical resources, distracting staff from their daily activities, and using office space. If a service provider were to allow all of its clients to pay audit visits independently and at times that suited them, it is not difficult to imagine that this could cause significant disruption to the provider's activities.

4.2 Third-Party (Service) Audit Procedures

Many service providers thus prefer the second method of allowing their clients to obtain comfort over internal controls: the service audit report. In this situation, the service provider itself mandates an external auditor to review its internal controls and provide an opinion on the effectiveness and efficiency of these controls [18]. This report is then made available to the service provider's customers who can incorporate its findings into their own evaluation of the control environment.

The advantages of such an approach for the service provider are clear. They meet their requirements to provide reliable information about their control environment, while avoiding the inefficiencies of having multiple audit teams visiting their premises and having to provide repeated briefings, explanations, and copies of documentation. If the audit partner chosen has a size and structure that allows consistency of team membership and a minimum of rotation, there will also be the advantage of familiarity year to year with business practices and documentation standards as applied by the service provider, which will also increase the efficiency of the audit process.

The report issued by the auditor can take many forms, but for many years the de facto international standard to be followed was the Statement on Auditing Standards No. 70: Service Organizations, commonly abbreviated as SAS 70.

This standard specifies two kinds of reports, named Type I and Type II. A Type I service auditor's report includes the service auditor's opinion on the fairness of the presentation of the description of controls that had been placed in operation by the service organization and on the suitability of the design of the controls to achieve the specified control objectives (thus corresponding to the concept of design effectiveness

as discussed above). A Type II service auditor's report includes the information contained in a Type I service auditor's report and includes the service auditor's opinion on whether the specific controls were operating effectively during the period under review. Because this opinion has to be supported by evidence of the operation of those controls, a Type II report therefore also includes a description of the service auditor's tests of operating effectiveness and the results of those tests.

SAS 70 was introduced in 1993 and effectively superseded in 2010 when the Auditing Standards Board of the American Institute of Certified Public Accountants (AICPA) restructured its guidance to service auditors, grouping it into Statements on Standards for Attestation Engagements (SSAE), and naming the new standard "Reporting on Controls at a Service Organization". The related guidance for User Auditors (that is, those auditors making use of service auditors' reports in the evaluation of the business practices or financial statements of organizations making use of the facilities provided by service providers) would remain in AU section 324 (codified location of SAS 70) but would be renamed Audit Considerations Relating to an Entity Using a Service Organization. The updated and restructured guidance for Service Auditors to the Statements on Standards for Attestation Engagements No. 16 (SSAE 16) was formally issued in June 2010 and became effective on 15 June 2011. SSAE 16 reports (also known as "SOC 1" reports) are produced in line with these standards, which retain the underlying principles and philosophy of the SAS 70 framework. One significant change is that management of the service organization must now provide a written assertion regarding the effectiveness of controls, which is now included in the final service auditor's report [19].

Internationally, the International Standard on Assurance Engagements (ISAE) No. 3402, *Assurance Reports on Controls at a Service Organization*, was issued in December 2009 by the International Auditing and Assurance Standards Board (IAASB), which is part of the International Federation of Accountants (IFAC). ISAE 3402 was developed to provide a first international assurance standard for allowing public accountants to issue a report for use by user organizations and their auditors (user auditors) on the controls at a service organization that are likely to impact or be a part of the user organization's system of internal control over financial reporting, and thus corresponds very closely to the old SAS 70 and the American SSAE 16. ISAE 3402 also became effective on 15 June, 2011.

The importance of understanding the related guidance for user auditors (which also applies, by extension, to user management) is critical. When a service audit report is received, the reader need to follow a number of careful steps in order to be certain that the report is both useful and valid before beginning to draw any conclusions from it.

These steps include:

1. Confirming that the report applies to the totality of the period in question. This is of particular importance when using a service audit report in the context of obtaining third-party audit comfort for a specific accounting period, but also applies to more general use of the report. If management's concern is over the effectiveness of internal controls for the period from 1 January to 31 December 2012, say, and the audit report covers the period from 1 April 2012 to 31 March 2013, how valid is it for management's purposes and how much use can they make of it? In the simplest of cases, a prior period report will also be available that will provide the

necessary coverage, but in other circumstances this may not be the case and management will have to turn to other methods to acquire the comfort that they need. These could include obtaining representations from the service manager that no changes had been made to the control environment during the period outside the coverage of the audit report and that no weaknesses in internal control effectiveness, either design or operational, had been identified during that period. Other methods could involve the performance and review of internal controls within the client organization, a subject to which we will return below.

2. Confirming that all the systems and environments of operational significance to the organization were included in the scope of the audit report. Management will need to have an understanding of the platforms and applications that are being used for their purposes at the service provider, including operating systems, databases, middleware, application systems, and network technologies, and be able to confirm that these were all appropriately evaluated. Very often in the case of large service providers a common control environment will exist, under which they apply identical internal control procedures to all of their environments: if the auditors have been able to confirm that this is the case, then it is not inappropriate for them to test the internal controls in operation around a sample of the operating environments, and management can accept the validity of their conclusions without needing the confirmation that their particular instance of the database, for example, had been tested. Should the coverage of the audit not meet management's requirements, again management would need to evaluate the size and significance of the gap and consider means by which they could obtain the missing assurance.
3. Understanding the results of the work done and the significance of any exceptions or weaknesses noted by the auditors. Generally speaking, if the work has been performed to appropriate standards and documented sufficiently, and if the conclusions drawn by the auditors are solidly based on the evidence, this step should be reasonably straightforward. Management should, however, guard against skipping to the conclusion and, if the report contains it, the service provider's management attestation, and blindly accepting the absence of negative conclusions as being sufficient for their purposes. Each weakness in the design or the operation of controls should be considered, both individually and cumulatively, to identify any possible causes for concern. This is because it is possible that weaknesses identified during the audit, considered to be insignificant within the overall framework of internal controls could potentially be significant in respect of the specific circumstances (use of systems and combination of technologies, for example) of one particular customer.

4.3 Other Procedures

Should such an independent audit report not be available, or only provide partial coverage of the control environment or the period in question, and should it in additional be impossible or impractical to the organization to perform their own audit procedures at the service provider's premises, a third way of obtaining comfort over operations is needed. This method can be summarized as consisting of two groups of activities: internal controls performed within the organization over the activities of the service provider, or audit procedures designed to evaluate the completeness, accuracy, and integrity of the service provider's processing and outputs.

First, internal controls can be designed that allow local management to monitor and evaluate the activities of a third party. These can take the form, for example, of procedures to track the responses of the service provider to requests for changes: if there is a process by which the customer asks the service provider to change access rights to an application or a datastore to correspond to the arrival or departure of a member of staff, management can track these change requests, and the responses of the service provider in order to ensure that the correct actions have been undertaken.

Very often the contract terms with the service provider will include regular meetings and reporting mechanisms through which the provider will present status updates, usually in the form of progress against KPIs and lists of open points, and management can ask questions and ensure that everything is under control. These meetings can form the central points of control activities for management, as a structured means of ensuring that they are monitoring the performance of the service provider in a regular and consistent way, observing long-term trends, and identifying anomalies.

Secondly, audit procedures can be designed along similar principles to the above, based on the expected results from the service provider's activities. An example of this would be extracting at the period end a list of user access rights to the organization's applications and comparing these to expectations, expectations based on management's understanding of the access requirements and on the instructions given during the year to create, modify, or delete access rights. If the rights correspond, this can provide a layer of assurance to management that both the service provider's internal procedures and controls are operational and that access to their applications and data is being appropriately managed.

With an appropriately selected range of internal controls and audit procedures, management can, therefore, obtain a certain level of comfort over the existence and quality of the control environment in place at the service provider.

4.4 Timescales and Logistical Concerns

It is important to recognize that the process of achieving the confidentiality, integrity, and availability of data is likely to be lengthy. Data growth is one of the most challenging tasks in IT and business management, with increasing quantities of data being generated in-house within organizations and being acquired from external sources. It is, therefore, essential to develop a structured plan for overall data management within with the steps necessary for data accumulation, security, and transfer can be carried out in a systematic and reliable manner.

In practical terms, management needs to take a number of operational decisions at an early stage in the planning process. Even before decisions can be taken about which data should be retained within an organization's infrastructure and which should be outsourced, more fundamental decisions need to be taken in respect of how, where, and by whom data will be cleaned, structured, collated, error checked, completed, or even clearly marked for deletion or a separate archiving process.

The "how" aspect will almost inevitably be addressed by a combination of automated tools and manual intervention. Software will be necessary be processing and transforming large quantities of data, but human intervention will be essential for defining and implementing parameters, reviewing output, and making ongoing decisions.

The "where" aspect throws up a question that anticipates, rather, the questions of data security that the move toward outsourcing and third-party storage also throws up.

Arguments could be made on the grounds of costs and logistics that the data preparation process should be performed offsite at a third-party site, on purpose-built infrastructure and away from the organization's internal networks. This would be in order, for example, to prevent excessive strain on resources caused by intensive processing and by large quantities of data passing across the network. Such a solution would, however, introduce the additional complication of ensuring adequate security over the data once the datasets have left the organization's security perimeter.

The "by whom" aspect will concern the use of internal resources, insofar as they can be spared, and external resources. Depending of the amount and the nature of the data to be processed, it may be appropriate to bring in resources from outside to perform aspects of the work. Internal resources will always be necessary, however, both from IT to provide technical input, and from the business side as users who understand where the data come from, what they represent, and how they relate to each other. A common failing in any data cleaning or migrating exercise is to view it as a purely technical procedure, whereas in practice the input from experienced and knowledgeable data owners and end-users is critical.

The timescale for such a project will depend on several factors, including the quantity and the nature of the data to be prepared, the availability of resources, and the priority set by senior management for the process. Experience of such projects would indicate that management should be setting their expectations in terms of months rather than weeks, however, and that if data quality and security are really expected, there is no scope for cutting corners.

5 Proposed Approach and Solutions

In order to gain an independent perspective on how the questions of cloud facilities and unstructured data were affecting organizations in Switzerland, the author performed a survey in 2011 and 2012.

5.1 Background to the Survey

50 questionnaires were sent out to contacts in 43 organizations across Switzerland, using the lead author's business experience to identify correspondents across a range of industries and sectors of activity who would be likely to respond. Completed questionnaires were received from 34 organizations, with three sending two responses and two sending three. The organizations that declined to respond did so on the grounds of confidentiality, not wishing to divulge details of their IT strategy or approach to security to a third party.

The organizations were selected in order to provide a wide cross-section of the range of IT environments and attitudes to the management of data and the use of new technologies. Of the organizations that did respond, five were publicly owned, eighteen privately, and the remaining eleven were public administrations or NGOs. The breakdown by organization size also shows variety: eight with less than fifty employees; ten with between fifty-one and one hundred; five with between one hundred and five hundred; and eleven with more than five hundred.

The rationale behind sending multiple questionnaires to the same organization was to attempt to discover whether there would be cases of poor communication of strategy or

developments within those organizations. It is the author's experience of large corporations in particular that there can be differences in the understanding of the overall strategy, the firm objectives, the initiatives undertaken, and the impact on users between top management, middle management, and IT management, for example.

5.2 The Survey

The questions in the survey were split into two sections, dealing with data security within the organization and with data storage in the cloud, as follows:

Table 1. Survey questions

Section A

Q1	Is there an overall policy concerning data management, security, and retention?
Q2	Is there awareness at the level of senior management and/or security management of the concept of "unstructured data"?
Q3	Has there been any assessment of whether the organization should be concerned, from security or efficiency perspectives, about the existence of unstructured data?
Q4	Has the organization established any guidelines on the classification of data according to their sensitivity, age, and relevance?
Q5	Has there been any structured attempt to identify and quantify the data stored around the organization outside centralized databases?
Q5A	If so, was this a manual process?
Q6	Does the organization have policies on the extraction, use, and storage of data by end-users?
Q6A	If so, is compliance with these policies monitored and enforced, and how?
Q7	Are there policies and/or restrictions in place of the use of removable storage media for file transfer or storage?
Q7A	If so, is compliance with these policies monitored and enforced and how?
Q8	Does the organization have any mechanisms for determining whether there have been breaches of security or confidentiality in respect of its sensitive data?

Section B

Q9	Is the organization using, or planning to use, cloud computing services for the storage of data?
Q10	If yes, have policies and guidelines been drawn up for the nature of data that can be stored in this way?
Q11	If cloud computing is being used, is the organization's approach based on the centralized management, monitoring, and retrieval of data, or do departments and/or individuals retain responsibility for their data?

5.3 Survey Results

Table 2. Survey results

Section A	Yes	No	Yes %	No %
Q1	32	9	78%	22%
Q2	22	19	54%	46%
Q3	17	24	41%	59%
Q4	9	32	22%	78%
Q5	6	35	15%	85%
Q5A	6	0	100%	0%
Q6	14	27	34%	66%
Q6A	5	9	36%	64%
Q7	12	29	29%	71%
Q7A	8	4	67%	33%
Q8	4	37	10%	90%
Section B				
Q9	39	2	95%	5%
Q10	6	33	15%	85%
Q11	7	32	18%	82%

5.4 Interpretation and Analysis of the Results

The results demonstrated two major trends in respect of unstructured data. The first was that although 78% of organizations reported having designed and implemented an overall policy concerning data management, security and retention, the exceptions being overwhelmingly small organizations with informal internal control structures, there was little systematic followup in terms of the management of unstructured data or in terms of managing and monitoring the use of data by users. 54% reported that senior management were aware of the issue of unstructured data; but only 41% reported that a risk analysis had been carried out to evaluate their exposure; 22% reported that guidelines for the classification of data had been developed and published; and only 15% reported having carried out a structured attempt to identify and quantify the data stored outside centralized databases. In each case it was a large multinational company that had undertaken such an initiative; interestingly, each one reported that the process had been largely manual.

From the perspectives of internal controls and good corporate management, many of these numbers are worryingly low. That more than three quarters of organizations report the existence of an overall policy in respect of data management is a reasonable starting point, but the lower numbers in respect of detailed data management indicate that few organizations had progressed further than the big picture, high-level aspects of data management at the time this survey was performed.

In respect of the management of user activities related to data handling, 34% reported having policies on the extraction, use, and storage of data by end users, and only 36% of these were able to report that compliance with these policies was monitored and enforced. Only 29% of organizations reported having policies or procedures in place concerning the use of removable storage media for file transfer or storage, while only 67% of this small subset were able to describe compliance mechanisms. Finally, only 10% of organizations were able to describe the existence of mechanisms for determining whether breaches of security or confidentiality in respect of sensitive data had occurred.

These rather low numbers suggest that organizations will have a significant amount of work to do in collating, cleaning, and preparing data. The responses suggest an overwhelming absence of effective internal control to date over data usage, and a lack of certainty over the nature, quality, and integrity of the datasets in use around organizations, factors that will automatically increase both the amount of scoping work that needs to be carried out and the quantity of detailed cleaning and tidying of data.

In the cases where more than one response was received from an organization, it was noted that end users were less aware of the existence of strategies and policies than senior management, a situation that is consistent with the author's long experience of auditing large organizations. A key difficulty in the implementation of internal controls within an organizations relates to ensuring that details of the control objectives and activities, their importance, their relevance, their nature and their follow-up, permeate sufficiently through the organization so that controls are operated effectively and efficiently, improving processes rather than hindering them, and with evidence of their performance and output being available for timely review and, if necessary, corrective actions.

In respect of the use of cloud computing facilities for the storage of data, 95% of the responses reported that the organization was using or was planning to use such facilities. The reasons most frequently given were cost management, flexibility, and a desire to streamline IT activities to concentrate on more value-added activities in-house. Of these organizations, however, only 15% had drawn up policies and guidelines concerning the nature of data that could be stored in this way, and only 18% were adopting an approach based on the centralized management, monitoring, and retrieval of data rather than leaving it to departments or individuals to manage.

Once again, the wide absence of overall policies and guidelines and the tendency as reported to adopt potentially uncoordinated approaches to project scope and management runs counter to established good practice in respect of internal controls. Without clear and enforced structure, inconsistency is likely to become a significant barrier to success. In addition, delegating down to departments or individuals increases the risks of decisions being taken without sufficient skills, experience, or perspective.

Overall it was possible to draw a clear distinction between large and small organizations and publicly and privately owned ones in respect of their approach to structured and formalized control environments. It was also possible to identify with high accuracy the responses received from publicly owned corporations and those from organizations subject to other strict and demanding controls requirements such as Sarbanes-Oxley or local industry or environment-specific regimes. It was also possible to identify organizations that had significant internal audit or internal controls functions, or that had been alerted to the risks involved in going through a process of discovery in the context of a legal dispute.

6 Summary Evaluations and Lessons Learned

The tentative conclusions that can be drawn from this very specific and targeted survey are the following.

First, the importance of having a structured and documented approach to data management and security is widely understood among the organizations surveyed, with a particularly positive attitude toward risk management and compliance from larger organizations and those subject to definite compliance regimes because of their ownership or industry. How this understanding actually translates into positive measures designed to ensure compliance is another question, however, and IT security managers in particular reported seeing greater enthusiasm for establishing policies within their organizations than for implementing and complying with the necessary procedures. There was also a question of priorities and resources raised by smaller organizations that did not feel that such policies corresponded to or were a part of their core daily activities.

Secondly, the concept of unstructured data and the particular challenges posed by such data is reasonably widely understood, but there has been little activity outside large publicly owned corporations to address the issue in any systematic way. In general IT departments had a good grasp of the nature and impact of the matter, and the

subject was frequently raised by internal and external auditors and by legal advisers, but it was rarely considered to be a subject of great priority by senior management.

Thirdly, even if thought has been given within some organizations to managing employee access to and extractions of sensitive data, in the majority of cases monitoring is weak and compliance cannot be ensured. Generally speaking, users who have access to data stored in centralized databases tend to have the ability and the opportunity, and frequently the encouragement, to extract those data and use them for analytical or reporting purposes. Once the data have been extracted, access controls around them are usually weaker, often being restricted to network or workstation access controls, and even these restrictions reach their limits once data are copied onto portable devices such as USB keys.

Fourthly, very few organizations are in a position of being able to detect reliably whether their security has been breached or their data compromised, even when all their systems and data are hosted and managed internally. Indeed, in respect of both security breaches and employee misuse of data, it was reported that incidents were typically identified either by chance or on the basis of information received, rather than on the basis of regular and reliable compliance measures. Of course, in practice being able to design, implement and monitor the operation of such control activities is frequently nontrivial and requires competence, resources, and careful planning. Whether organizations would be able to apply such controls to outsourced data, or be satisfied by the monitoring and reporting services provided by their cloud service provider, is the same question with an added layer of complexity.

Fifthly, the idea of cloud computing as a financially attractive option for outsourcing a number of traditional IT activities including data storage is widespread and there is a great deal of enthusiasm for it across a wide range of organizations, but this enthusiasm is not yet being widely and systematically backed up by detailed risk assessments and careful consideration of the approaches needed to identify, classify, manage, and monitor the data being transferred into the cloud.

The comments provided alongside the answers also provided useful information. In particular, several respondents referred to the different types of cloud that are beginning to exist: in certain industries such as financial services it would be unthinkable to use cloud services in which confidential data might be stored outside Switzerland or for which it would be difficult to obtain adequate audit comfort over key concerns, but the use of some kind of industry-specific Swiss cloud, perhaps setup as a joint venture, with appropriate controls and safeguards in place, might be conceivable.

Several respondents also flagged up the importance of the proper management of backup media, which of course need to be subject to the same policies and procedures for data management and security as live datasets. From the perspective of the management who will retain the responsibility for ensuring the availability and reliability of system and data backups for good practice and going concern reasons, and for the auditors who will be verifying this, it will be a challenge to identify exactly which data are backed up where, how this is managed from a security and availability perspective, and what the timelines, sequences, and interdependencies would be for restoring part or all of a missing dataset.

Within all businesses there is constant pressure to reduce costs and cloud computing could be seen as an effective method of managing and reducing costs, particularly in the short term. If there are no significant in-house IT systems, a case will always be made for reducing to a bare minimum, or even eliminating entirely, the IT function, thereby, reducing staff costs alongside the operational costs of monitoring and maintaining systems.

The decision to choose cloud computing services is not one to be taken lightly. For individuals, the use of personal services in the cloud (such as Facebook, gmail, and Dropbox) is, or rather should be, a matter of a calculated assessment of risks and benefits, for the potential negative impacts of breaches in security, for example, can be significant. For businesses and other organizations, the same concerns apply but on a larger scale. For all organizations that have a responsibility to keep their, and others', data secure and confidential, the decision can only be taken after a detailed analysis of how they will obtain, and continue to obtain, the necessary comfort that this is the case. If they cannot build into their own procedures, into enforceable contract terms, and into audit plans, the means of confirming the confidentiality, integrity, and availability of their systems and data, they should not consider externalizing it.

In addition, organizations should not forget the immediate costs and efforts involved in moving onto the cloud. Datastores need to be identified, classified, cleaned up, and archived, and serious technical and operational decisions are needed to determine what data will sit where. This will frequently be a project of a significant size requiring expertise, resources and input from a number of people across the business who understand the business, the systems, the data, and their use.

Experience of traditional outsourcing suggests that it is very easy for organizations to overestimate the cost savings generated by a move toward service providers and to underestimate the amount of internal competence and dedicated management required to make a success of such initiatives. As long as the organization relies on its data and retains responsibility for all aspects of its business from a regulatory perspective, it will need to ensure that its management of the relationship with its service providers and its access to critical operational information are both adequate and appropriate. Typically this will require retaining or recruiting skilled, experienced, and reasonably senior staff to liaise with and monitor the performance of the service provider.

The long-term consequences of opting for a cloud solution also need to be examined. Once on the cloud, systems and data are likely to stay there, and feasibly with the same provider. What begins as a simple and cost-effective solution to a small problem could develop into a long-term strategic commitment with little scope for alteration.

7 Further Research Issues

It will be necessary to monitor the development of the use of cloud services for data storage from a controls perspective in order to see how such use develops, whether it becomes widespread and whether it does become a factor in assessing the quality of internal controls. Perhaps practical and effective operating procedures will be developed and become standard very rapidly, so that the subject does not become a major

concern for controls managers, or perhaps the take up of the technologies will not be great, because of perceived controls issues or questions of cost or access.

It will also be interesting to study the impact of the uptake of such services on audit opinions and compliance reports. This will surely be a major driver in the development and the use of these services: if organizations find themselves subject to adverse comments from regulators or external auditors, that will necessarily cause a slow-down in adoption. On the other hand, if clean reports are issued and no concerns are raised, take up will only be encouraged.

8 Conclusion

In common with all IT-related projects, initiatives in respect of data collation, and accumulation and in respect of data migration and transfer require a great deal of planning, strategic awareness and effective controls in order to ensure that the key security objectives defined by the organization continue to be met. Both managing unstructured data and managing the migration onto, and ongoing monitoring of, cloud services, present countless opportunities for the loss of the confidentiality, integrity and availability of data, to cite once again just the three most famous security objectives.

The use of large datasets for competitive advantage is highly tempting for many organizations from a tactical perspective, while the use of cloud services is attractive for a number of reasons, financial, operational, and strategic. The senior management of organizations tempted by such initiatives should be aware, however, that neither type of project can be successfully completed overnight, and that they should be prepared to provide the necessary resources, guidance, oversight, and supervision to ensure that the advantages obtained through the initiatives are not outweighed by decreased security, increased costs, or reduced comfort from internal controls.

References

[1] Marr, B.: Is This the Ultimate Definition of “Big Data” ?, http://smartdatacollective.com/node/128486 (accessed September 2, 2013)

[2] Bojilov, M.: Big Data Defined. In: ISACA Now (2013), http://www.isaca.org/Knowledge-Center/Blog/Lists/Posts/Post.aspx?ID=299 (accessed September 2, 2013)

[3] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A.: Big data: the next frontier for innovation, competition and productivity. McKinsey Global Institute, Washington DC (2011)

[4] Blumberg, R., Atre, S.: The Problem with Unstructured Data. DM Review (2003)

[5] National Institute of Standards and Technology. The NIST Definition of Cloud Computing. Special Publication 800-145, NIST, Gaithersburg (2011)

[6] Committee of Sponsoring Organizations of the Treadway Commission. Guidance on Monitoring Internal Control Systems. AICPA, New York (2009)

[7] Krutz, R., Vines, D.: Cloud Security: A Comprehensive Guide to Secure Cloud Computing. Wiley Publishing Inc., Hoboken (2010)
[8] Internal Standards Organization. ISO/IEC 27001 Information Technology - Security Techniques – Information Security Management Systems. ISO/IEC, Geneva (2005)
[9] National Institute of Standards and Technology, Information Security. Special Publication 800-100. NIST, Gaithersburg (2006)
[10] Parker, D.: Toward a New Framework for Information Security. In: Bosworth, S., Kabay, M. (eds.) Computer Security Handbook, 4th edn. John Wiley & Sons, New York (2002)
[11] Organisation for Economic Co-operation and Development (2002) OECD Guidelines for the Security of Information Systems and Networks: Towards a Culture of Security, http://www.oecd.org/sti/ieconomy/15582260.pdf (accessed January 8, 2014)
[12] Adolph, M.: Distributed Computing: Utilities, Grids and Clouds. ITU-T Technology Watch Report 9, ITU, Geneva (2009)
[13] Gantz, J., Reinsel, D.: Extracting Value from Chaos. IDC iView (2011)
[14] Information Systems Audit and Control Association, Big Data: Impacts and Benefits. ISACA, Chicago (2013)
[15] Bittman, T.: A Better Cloud Computing Analogy. Gartner Blogs (2009), http://blogs.gartner.com/thomas_bittman/2009/09/22/a-better-cloud-computing-analogy/ (accessed January 8, 2014)
[16] Information Systems Audit and Control Association, Security Considerations for Cloud Computing. ISACA, Chicago (2012)
[17] Ghernaouti-Hélie, S., Tashi, I., Simms, D.: Optimizing security efficiency through effective risk management. Paper presented at the 25th IEEE International Conference on Advanced Information Networking and Applications (AINA 2011), Biopolis, Singapore, March 22-25 (2011)
[18] American Institute of Certified Public Accountants, Quick Reference Guide to Service Organizations: Control Reports. AICPA, New York (2012)
[19] American Institute of Certified Public Accountants, Service Organizations: Reporting on Controls at a Service Organization Relevant to User Entities' Internal Control over Financial Reporting Guide. AICPA, New York (2011)