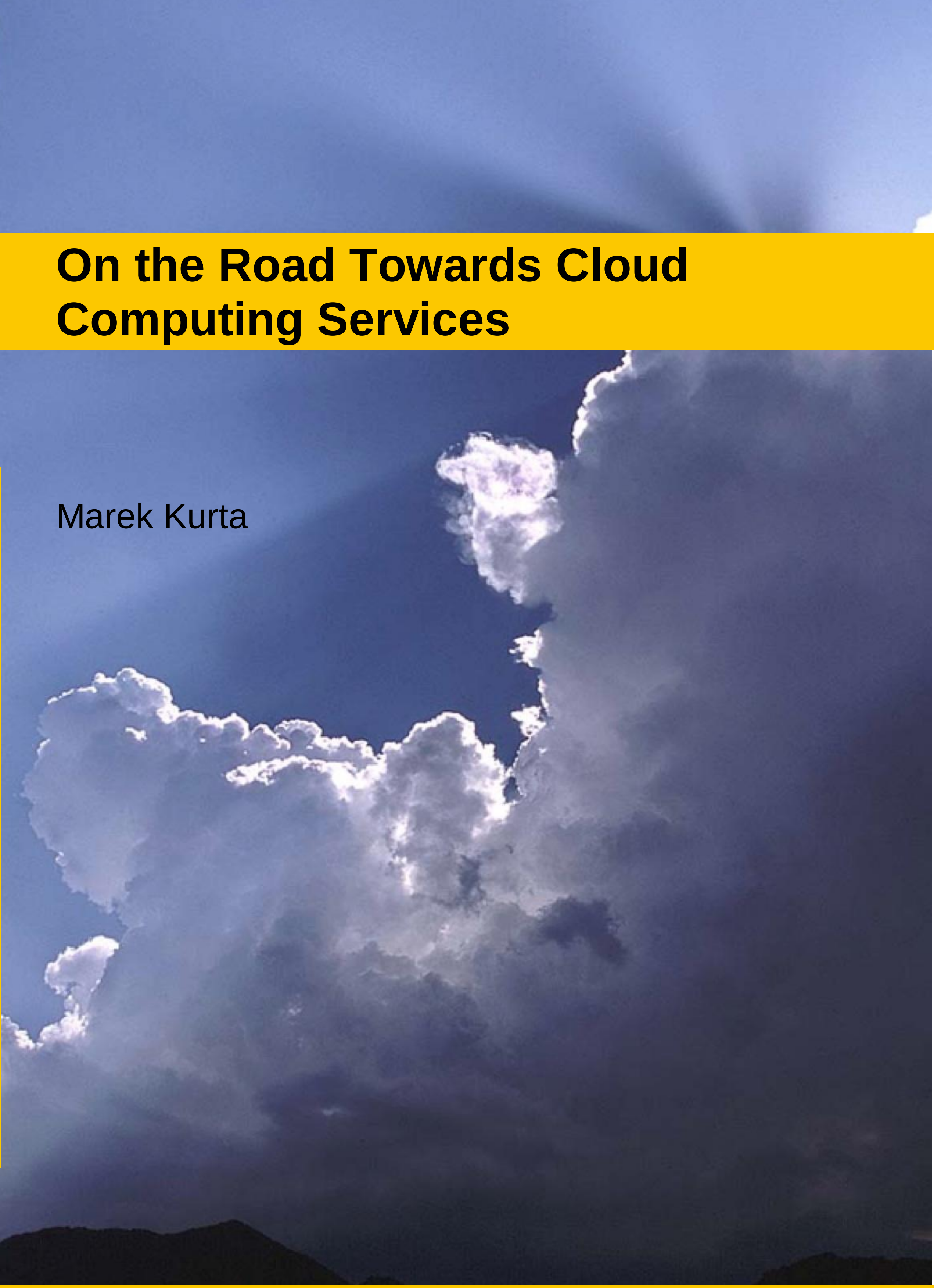
On the Road Towards Cloud Computing Services



Marek Kurta

Erasmus School of Economics

Master Thesis Economics & Informatics

Economics & ICT program

Student ID: 333055

Supervisor: Ad de Visser

Second Supervisor: Bram Borgman

ERASMUS UNIVERSITEIT ROTTERDAM

August 2010

Acknowledgements

I would like to thank my supervisor Mr. Ad de Visser for helping me throughout the process of writing this thesis. He always had useful remarks and comments, and always helped me to stay on track. I am also grateful for the help from my co-supervisor Mr. Bram Borgman, who gave me additional ideas on how to improve this work.

I would also like to use this opportunity to thank the respondents of the interviews and surveys that were conducted for this thesis. By sharing their experience with the discussed topics, they have greatly contributed to the final outcome of this thesis.

Last but not least, I would like to thank my family, who made my studies abroad possible. Without their support, I would not be where I am now.

NON-PLAGIARISM STATEMENT

By submitting this thesis the author declares to have written this thesis completely by himself, and not to have used sources or resources other than the ones mentioned. All sources used, quotes and citations that were literally taken from publications, or that were in close accordance with the meaning of those publications, are indicated as such.

Abstract

This thesis is a descriptive study of the prerequisites of companies considering implementing cloud computing technology. A description of the technical possibilities is presented and the topic of cloud computing implementation is approached from a user perspective. Firstly, it elaborates on security, compliance, legal and other issues of public cloud use. Secondly, it attempts to describe the main roadblocks slowing the adoption of cloud computing services. This research concludes that in order to overcome these issues, a company is required to implement a data classification system. For this reason, this topic is one of the central themes of the thesis, whereas finally a system that should automate the whole process of data classification is outlined.

Keywords:

Cloud computing, implementation, prerequisites, automatic data classification.

List of Figures

List of Tables

Table of Contents

1 : Introduction

1.1 Introduction

Individual users in general do not have a problem with posting their personal data on the internet. Actually, sometimes they even want this kind of data to be shared. They upload pictures, videos and write blogs. Companies, on the other hand are much more cautious. They seek assurance that their confidential data will be secure, even if allowed access through internet.

In the last couple of years a new information technology (IT) paradigm called cloud computing emerged on the internet. It is promising to increase IT efficiency and on the other hand reduce IT costs considerably, making it a very attractive technology. Due to the recent trend and pressures to cut costs, businesses are considering it, if not using it already.

Because of the nature of the cloud computing concept, its immaturity and the absence of standardization, companies have legitimate worries on their data security when considering using its services.

In this master thesis, we will look at some of the prerequisites of companies' information systems (IS) security in order to begin using cloud computing services. Moreover, the research will show that automatic data classification is a necessary condition for a successful full implementation. Therefore we will elaborate on the challenge of automatic data classification and its impact on the current IS. Besides answering the research question, this thesis will look at cloud computing security from a user perspective and will deal with its benefits, costs and requirements for installation.

1.2 Research questions

RQ: ***What are the current roadblocks for companies considering cloud computing services?***

SQ1: What technical possibilities do cloud computing services offer?

SQ2: What are the security concerns of companies' information systems?

SQ3: What are the possible security measures to address the security concerns?

SQ4: How necessary is automatic data classification when considering using cloud computing services?

SQ5: How to classify data automatically?

1.3 Theoretical background

Cloud computing is a relatively new term that describes a concept which builds up naturally from previous advancements in computing technology. The concept is basically about offering computing as a service accessible on-demand over a network. Since there are many definitions of cloud computing in the literature, it shows that the concept is still immature and evolving. Various definitions have similar elements though; therefore according to their research, L. Vaquero et al. (2009) propose the following one:

Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services). These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized service level agreements.

Service providers have to ensure that they can be flexible with their service delivery while keeping the users isolated from the underlying infrastructure (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009). As the proposed definition of cloud computing suggests, services of cloud computing are offered through virtualized resources, which are referred to as virtual machines (VM). A VM abstracts a physical machine with the use of software and offers great flexibility to users together with other advantages. Main attributes of VM are software compatibility, isolation, encapsulation and performance (Rosenblum, 2004).

There are three scenarios of cloud computing services: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). These concepts are also collectively referred to as service-oriented architecture (SOA) (Vouk, 2008). IaaS can offer a full virtual infrastructure including storage and processing capacity which is easily accessible over the internet. PaaS offers an environment with development tools to add, change and host application software. The last scenario of SaaS offers the users access to software for which they do not need to own licenses.

The paradigm of cloud computing is not a new concept. It roots back to the year 1969 when the creators of ARPANET (Advanced Research Projects Agency Network) envisioned the spread of “computer utilities” which would service individuals in homes and offices like for example electricity does today (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009). The creation of the internet was the primary milestone in achieving this goal. The emergence of Web 2.0 and the widespread of broadband internet paved the way for cloud computing which combines features of its predecessors. As the technology advances, so do the different paradigms promising to deliver IT as services. The predecessors of cloud computing are for example Grid computing, Peer-to-Peer computing, Services computing and Market-oriented computing[1].

Individual internet users are used to cloud computing services in the form of SaaS. The most recognized examples are e-mail services and other web applications like the ones from Google for example. Social networks are also good examples of cloud computing. These applications are accepted well in general, and users do not have problems with storing their personal information in these clouds, which is proved by the number of active users of Gmail or Facebook. Businesses on the other hand have serious security concerns when it comes to using cloud services. Most of these concerns are related to the diminishing information system boundaries, while others are related to the availability of the service. This creates various business process risks which need to be tackled.

Security is usually referred to as a combination of three factors: confidentiality, integrity and availability. What is more, cloud computing raises three more factors, namely compliance, policy and risk (Hobson, 2009)[2]. For companies, sending confidential data over the internet and storing them in a cloud which is shared with other users is a problem. They need a guarantee that no one else will have access to their data and that integrity will be maintained. Since it is usually possible to access the cloud only by internet, availability is threatened in case of a connection outage. Companies are required to have alternative connections to the internet at their disposal at all times to ensure availability of access to the services. There are also problems with compliance requirements. For example, companies that handle their customers' credit card numbers need to comply with the PCI standard. There are also sector-specific regulations such as HIPPA for healthcare or FISMA and ITAR for organizations in the public sector. Naturally, there can be a mixture of regulations that need to be satisfied. Ensuring that a company's services will be managed correctly when they do not own the infrastructure of the data center is therefore a challenge. A company's data storage policy must not only address these compliance issues, but there are also many legal ones. The physical location of the data center that holds the data may be important. Local laws may threaten their security, since laws like the US Patriot Act allow the government to have virtually limitless access to any kind of information. On the other hand, the EU is in favor of much more strict protection of privacy. The last security factor of cloud computing is the risk of the service provider to stop doing business. In that case, the users need to know if it would be possible to transfer all their data and how long would this take. It is clear that even if the service provider can give reasonable assurance for confidentiality, integrity and availability, due to compliance, data storage policy and risk of service discontinuity, not all data is suitable for the cloud.

[1] Differences between these paradigms are summarized in chapter 2.2.3.

[2] A more thorough elaboration on this opinion is available in chapter 2.3.3.

An IS boundary also known as a perimeter contains all of the information assets in an IS (Raval & Fichadia, 2007). The problem of controlling an IS boundary is not new, although it is still a challenge. Cloud computing falls within the group of distributed systems. This means that instead of inventing completely new security solutions, we can learn from previous studies of this kind of system, even though cloud computing is a relatively new concept. Distributed computing technologies follow a similar pattern of interaction, where disparate and sometimes heterogeneous systems interact with one another over a common communication platform (Belapurkar, Chakrabarti, Ponnapalli, Varadarajan, Padmanabhuni, & Sundarrajan, 2009). The more gateways to a company's IS there are, the higher the risk of a security breach. Since the security measures at the system boundary cannot bring sufficient assurance, it is necessary to go deeper into the IS. We need to make sure that the company data in the cloud is secure as well as in the internal network. Consequently, we have to address security at the data level. Data flowing in a network needs to have incorporated information on how it should be treated. Companies have various policies for handling this kind of security and the most common solutions are encryption, access controls, rights management and auditing. The policies do not approach all the data in the same way, and define different levels of confidentiality. It would be too costly and inconvenient to treat all data as most confidential due to the tradeoff between security, usability and costs (Raval & Fichadia, 2007). In that case, using cloud services would be impossible, moreover the IS itself would be probably isolated from the internet completely.

1.4 Motivation

To be able to determine how to treat individual data and to see if it can leave the IS boundary and be sent to the cloud, a company should adopt data value management based on their policy. In cloud computing, perimeter security has no meaning. Extending this perimeter with the cloud to create a virtual boundary and apply security there is not enough for all types of data, simply because of some regulations mentioned in the previous section. One of the tools for this data value management is data classification. The cloud being a dynamic environment requires a dynamic and seamless integration with a company's IS, free of any manual intervention. For this reason, automation of the process is preferred.

Several researches have been done to justify why cloud computing can help companies to save considerable amounts of money on their IT, therefore the potential benefits of cloud computing are well understood. There were also researches dealing with security from the service provider's point of view. Others were dealing with security in communication between the users and the cloud, including the assurances they should seek from the service providers (Bean, 2009), (Parrilli, 2010).

Only a little is being said about the prerequisites of companies to enter this environment. The goal of this master thesis is to discover and examine these prerequisites of this cloud technology implementation in a form of a descriptive study. Studying these factors is important, because if a company does not have sufficient internal security policies and practices, introducing cloud computing technology only stacks new risks on top of the old ones. Consequently, this can result in a very expensive IT investment failure.

1.5 Methodology

For this master thesis, a literature review is conducted based upon various scientific articles, books, but also internet blogs to compensate for the shortage of scientific research. To better understand the issues of cloud computing integration within a company, well described predecessors of cloud computing are analyzed in books on distributed systems like for example by Kshemkalyani & Singhal (2008) and Belapurkar et al. (2009). In these books, problems and solutions for systems like grid computing are presented, but they lack an extension to cloud computing solutions. Generally it is possible to find descriptions of new problems with compliance and policies, but the literature does not present many concrete solutions. In order to be able to extend the model of systems like grid computing, a clear understanding of differences with cloud computing is necessary. For this reason, comparisons like the ones in a paper by Buyya et al. (2009) are observed. Information about cloud computing is also gathered from recent books like "Cloud Computing: A Practical Approach" (Velte, Velte, & Elsenpeter, 2010) or "Grid and Cloud Computing: A Business Perspective on Technology and Applications" (Stanoevska-Slabeva, Wozniak, & Ristol, 2010).

Due to the immaturity of cloud computing, this master thesis is written as a descriptive study to gather knowledge about the factors affecting the implementation of cloud computing. Moreover, a special emphasis is given to a concrete topic, namely automatic data classification. Information about data classification comes mainly from papers, reports and internet blogs written by security experts but also books like "Machine Learning and Data Mining for Computer Security" (Maloof, 2006) and "Machine Learning in Document Analysis and Recognition" (Marinai & Fujisawa, 2008) are examined, which are basically collections of scientific papers.

An empirical study is conducted with the goal to verify the collected knowledge and discovered issues slowing down the adoption of cloud computing technology. The empirical data is gathered based upon conducted interviews with companies that provide cloud computing services and also a survey has been made within companies which have recently implemented them. This thesis is divided into chapters, whereas each one talks of a different subject, but they are all interconnected and collectively follow a logical path.

1.6 Empirical data

Two kinds of sources are used for the empirical study. First of all are service providers who were asked open questions to encourage them in saying their personal opinions about the prerequisites of entering the cloud. Key people from Citrix Systems and IBM Slovakia were willing to provide their answers.

Main focus however, is given to the survey which concentrates on the implementation process itself. For this we could not have asked the service providers, because we assumed they would not point out any relevant problems. Instead, a selection of companies that are known to have recently implemented cloud computing services were approached. Some were willing to share their knowledge.

The limitation of the empirical study is the lack of a sufficient number of sources. Although surveys suggest that many companies are considering using cloud computing services, it is difficult to find a list of them who have, and get in touch with concrete people who were involved in these projects.

2 : Cloud Computing

2.1 Introduction

In this chapter we will look at an overview of the complexity of cloud computing. At first we will define the components and services that are being offered and then we will examine several factors companies need to consider carefully before entering the cloud environment. We will look at cloud computing from a cost and benefit point of view, but we will also briefly discuss security risks associated with combining the cloud with a company's IS.

2.2 Definitions

Any application needs a model of computation, a model of storage and, assuming the application is even trivially distributed, a model of communication (Armbrust, et al., 2009). As we have described in the first chapter, cloud computing represents a distributed system in which the IT infrastructure is offered as a service. The services can range from a single application to a whole infrastructure, including virtual computers with various operating systems (OS) and storage space. It involves a service oriented architecture, reduced information technology overhead for the end-user, great flexibility, reduced total cost of ownership, on-demand services and many other things (Vouk, 2008). These cloud services are in most cases accessed through an internet connection; therefore the term "cloud" is used, as it is how internet is depicted in network diagrams.

2.2.1 Components

There are three major components of cloud computing (Figure 1):

- Clients
- Datacenters
- Distributed servers

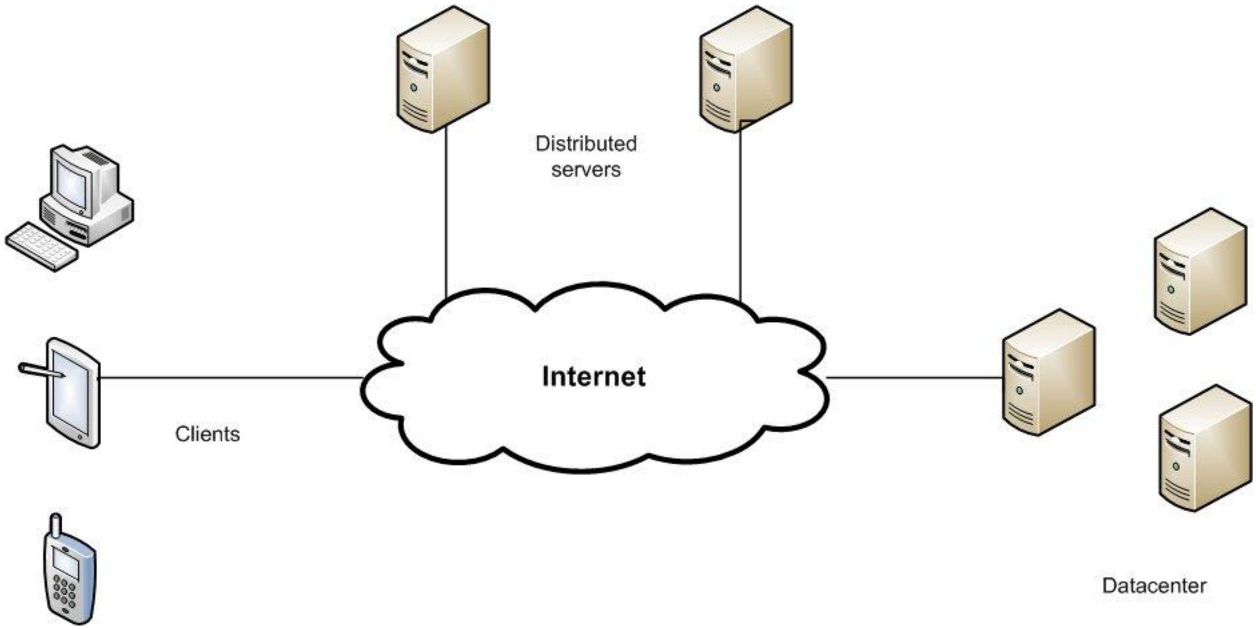


Figure 1: Components of cloud computing. Source: (Velte, Velte, & Elsenpeter, 2010).

2.2.1.1 Clients

Clients are the end users of cloud services. They can be further divided into thick, thin and mobile clients. The most common are thick clients which are normal standalone computers connected to the network, while thin clients are computers that only have internal memory, because they access all the data from the servers remotely. Mobile clients can be laptops, PDAs or various types of Smartphones. It is necessary to differentiate between the types of clients, because they propose various security challenges which we will describe in more detail later on in section 2.3.3.1.

2.2.1.2 Datacenters

Datacenters are large rooms with physical machines that provide storage and processing power. They house the applications that are accessed as services. Because of the decreasing prices of computers, datacenters of cloud computing providers are usually composed of hundreds to thousands of commoditized computers with specialized processors that have hardware support for virtualization.

2.2.1.3 Distributed servers

The servers to which the clients connect appear as if they are located in one datacenter, but in reality they are usually geographically distributed over various locations. Generally, the clients do not need to care about the complexity of the infrastructure that hosts their services. This is usually hidden from the client, but in some cases the geographical location of the server is important, which will be discussed later on in this chapter in section 2.3.3.2.

2.2.2 Services

According to the complexity of offered services, we generally define three types of concepts:

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)

It is possible to find different names for these concepts in literature (Schuller, 2008) and especially between the vendors[3], and naturally there are different ones consisting of combinations of these. Attempts have been made to track all the different "as-a-Service" acronyms (Laird & Oracle, 2008), but they are often too complicated, blurring the lines between the paradigms and making them misleading. However, for an overall understanding of the possibilities cloud computing can offer, it is enough to understand how these three concepts differentiate.

2.2.2.1 Software as a Service

SaaS is a very popular service in which suppliers deliver web applications to its users. The pioneer of providing SaaS was a company called Salesforce.com. Another common example of a SaaS provider is Google with its email and office tools like word processor, spreadsheet or calendar. It could be any type of application and users can even create their own, hosting them on the provider's servers. They work well in the internet environment, as they are specially designed for it.

Microsoft has taken a different approach, by creating a service which they call Software plus Service. These services are commercially known as services with the words Live or Online attached to them. They enable a seamless integration of their desktop applications with cloud services. This allows for the comfort of having all the functionality of desktop software and the option to store files on-site, together with easy and convenient team collaboration functionalities. The newest Office 2010 natively supports some of these functions.

2.2.2.2 Platform as a Service

In order to avoid confusion of this service with Saas, it is good to imagine it as a cloud OS. The providers of the service enable its users to install their applications on a platform, which can provide any operating system or even emulate various types of hardware. With SaaS, the providers enable access to their own applications which they host on their servers, but with PaaS the users have greater possibilities for deployment of their own software which does not have to be specialized for the web browser. The processing power is scalable, so they do not have to worry about the number of servers they would have to use in-house, especially if the required power is only seasonal, for example when closing a financial year.

[3] For a figure of all major vendors in May 2009, please refer to (Laird, Cloud Computing Taxonomy at Interop Las Vegas, May 2009).

2.2.2.3 Infrastructure as a Service

This service can provide the functionalities of a whole infrastructure including storage, any platform and any number of desktops. Therefore, it is possible to find another name for this service, which is Everything as a Service. The current leaders in cloud computing are Google and Amazon (Velte, Velte, & Elsenpeter, 2010). There can be also variations of IaaS, where one is referred to as Hardware as a Service. In contrast to SaaS or PaaS which provide software, Hardware as a Service only provides for virtual hardware on which a user can install anything. In theory, it would be sufficient for a company to have only thin clients in-house and host everything in the cloud. This is not realistic at this moment though, due to the fact that the pricing of cloud services may not be beneficial for all sorts of businesses, but especially due to many security concerns which we will discuss throughout this thesis.

Cloud computing is not a service that can suit every type of organization and requires a thorough consideration of many factors. In order to mitigate some of the security concerns, some companies decide to create so called private clouds. This means that they are the only company occupying it and they probably run the datacenter in-house. This creates a more controlled secure environment, but it comes at a cost of losing some major benefits of public clouds, for example the lack of need to invest heavily into the infrastructure[4]. Because we are rather interested in the more challenging environment of the public clouds, when we mention cloud computing throughout this thesis, we will generally refer to public cloud computing.

2.2.3 Differentiation from other paradigms

A look into the predecessors of cloud computing reveals that distributed systems have been used for a long time. Therefore, the security risks of communication between the client and the service provider are well understood and most have a certain degree of solutions.

Cloud computing is often confused with grid computing although the concept is different. In grid computing, a network of computers share their processing power to work on a single problem, while in cloud computing a series of smaller applications run on a system independently. Grid computing is often used for scientific and research purposes where a large computational power for a single task is required.

Because of the similarity with cloud computing, an analysis of its predecessors can be helpful to better realize various security threats related to accessing external hardware. Table 2 summarizes the differences between cloud computing, grid computing and cluster systems, which are

[4] A more thorough description of the benefits is described in section 2.3.1.

predecessors of the cloud. Other examples of SOA paradigms are Peer-to-Peer computing, Services computing, Market-oriented computing and Utility computing (Table 1).

Paradigm	Characteristics	Goal
Grid computing	Enables the sharing, selection and aggregation of a wide variety of geographically distributed resources owned by different organizations.	Solve large-scale resource-intensive problems in science, engineering and commerce.
Peer-to-Peer computing	Allows peer nodes (computers) to share content directly with one another in a decentralized manner without the notion of clients or servers.	Cost sharing or reduction, resource aggregation and interoperability, improved scalability and reliability, increased autonomy, anonymity or privacy, dynamism, and ad-hoc communication and collaboration.
Services computing	Focuses on the linkage between business processes and IT services so that business processes can be seamlessly automated using IT services (technologies include Service-Oriented Architecture and Web Services).	The SOA facilitates interoperable services between distributed systems to communicate and exchange data with one another, while Web Services provide the capability for self-contained business functions to operate over the internet.
Market-oriented computing	Views computing resources in economic terms such that resource users will need to pay resource providers for utilizing the computing resources.	Offer the incentive for resource providers to contribute their resources for others to use and profit from it.

Source: (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009), Own compilation.

Table 1: A comparison of cloud computing predecessors.

These are concepts that have been used for a longer time than cloud computing, therefore the same solutions to some threats could be extended to this system as well. Concrete threats that emerged in these distributed systems are listed in section 2.3.3.1.

Characteristics	**Systems**		
	Clusters	Grids	Clouds
Population	Commodity computers	High-end computers (servers, clusters)	Commodity computers and high-end servers and network attached storage
Size/scalability	100s	1000s	100s to 1000s
Node Operating System (OS)	One of the standard OSs (Linux, Windows)	Any standard OS (dominated by Unix)	A hypervisor (VM) on which multiple OSs run
Ownership	Single	Multiple	Single
Interconnection network | speed	Dedicated, high-end with low latency and high bandwidth	Mostly Internet with high latency and low bandwidth	Dedicated, high-end with low latency and high bandwidth
Security/privacy	Traditional login/password-based. Medium level of privacy – depends on user privileges.	Public/private key pair based authentication and mapping a user to an account. Limited support for privacy.	Each user/application is provided with a virtual machine. High security/privacy is guaranteed. Support for setting per-file access control list (ACL).
Service negotiation	Limited	Yes, SLA based	Yes, SLA based
User management	Centralized	Decentralized and also virtual organization (VO)-based	Centralized or can be delegated to third party
Resource management	Centralized	Distributed	Centralized/Distributed
Allocation/scheduling	Centralized	Decentralized	Both centralized/decentralized
Standards/inter-operability	Virtual Interface Architecture (VIA)-based	Some Open Grid Forum standards	Web Services (SOAP and REST)
Capacity	Stable and guaranteed	Varies, but high	Provisioned on demand
Failure management (Self-healing)	Limited (often failed tasks/applications are restarted).	Limited (often failed tasks/applications are restarted).	Strong support for failover and content replication. VMs can be easily migrated from one node to other.
Pricing of services	Limited, not open market	Dominated by public good or privately assigned	Utility pricing, discounted for larger customers
Internetworking	Multi-clustering within an organization	Limited adoption, but being explored through research efforts such as Gridbus InterGrid	High potential, third party solution providers can loosely tie together services of different Clouds
Application drivers	Science, business, enterprise computing, data centers	Collaborative scientific and high throughput computing applications	Dynamically provisioned legacy and web applications, Content delivery
Potential for building 3rd party or value-added solutions	Limited due to rigid architecture	Limited due to strong orientation for scientific computing	High potential — can create new services by dynamically provisioning of compute, storage, and application services and offer as their own isolated or composite Cloud services to users

Source: (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009)

Table 2: Key characteristics of Clusters, Grids, and Cloud systems.

2.3 The dilemma of using cloud computing in organizations

Several factors have to be clearly evaluated when considering using cloud computing services. The most obvious are the benefits, in order to determine if it is the right solution; costs, to balance them with the benefits and find out if the solution is worth it; but it is also important to see the status of security and its solutions to compare them with the risk appetite of the company.

2.3.1 Benefits

Probably the greatest benefit of cloud computing is its scalability of resources. It is ideal for a company that has volatile demand for computational power, or needs extra power temporarily or seasonally. If the user requires extra servers to perform a task, they are assigned to him automatically by the service provider in a relatively short time. On the other hand, the elasticity of the service enables to scale down immediately when the resources are not needed anymore. Moreover, the risk of estimating the expected workload is transferred to the service provider. All this is available through a pay-as-you-go scheme. There is no extra charge for the scalability, as running one server in the cloud for a thousand hours costs about as much as running a thousand servers in the cloud for one hour (Armbrust, et al., 2009).

Data mining is a great example of use for cloud computing through batch processing tasks. Data mining is sometimes used to better understand the customers, their behavior and to use these analytics to improve customer satisfaction and consequently increase sales. Although these analytics are very computationally intensive because they have to scan through a vast amount of data, the costs of these tasks might be comparable to making them in-house, although the speed at which they can deliver results is difficult to beat.

The speed at which a user can get a certain level of processing power is a great benefit related to the scalability of cloud computing services. This cannot be matched to regular in-house capabilities. If a company wants to run a project with a higher demand for power, they would have to spend hours or days preparing the hardware and installing it. This entire burden is shifted to the service provider and extra servers are available within minutes.

In case of IaaS and especially SaaS, the company should perform an analysis of how often certain employees need or use specific software. For example if there is an office suite which is used rarely on a computer that has it installed anyway, it might be shifted to a SaaS to save money on unnecessary licenses. Furthermore, computers that are not being used all the time could be transformed to thin clients, in which case the company would not only save money on software licenses, but also expensive hardware. The more thin clients a company would use, the better audit of user activity they could do. It would become very easy to track user behavior within the system, so machine learning and data mining techniques would be more successful in intrusion and insider threat detection (Maloof, 2006).

In case of using cloud services as storage, the scalability could be beneficial for backing up data for reasons like disaster recovery. In case the data is intended for long term storage and therefore is reasonably well encrypted before it is sent to the cloud, it might be even safer than kept in-house. It would be stored offsite with high security guarantees, and even if a malicious user or some government agency manages to get hold of it, it would be unreadable to them. On the other hand, not all data on a company's computer is confidential and often is redundant, stored on computers across the organization. In this case, centralization of the storage might save some costs, but it might also enhance team collaboration, as users might be able to access work even from the field or their homes.

Cloud computing is also excellent for start-up companies because they can enjoy high computational power without the need to invest heavily in hardware. They are then able to scale the demand as they grow. This increases competitiveness within their business.

2.3.2 Costs

Before a company closes a deal with a cloud computing provider, they negotiate several agreements. These are called service level agreements (SLA) in which the parties have to agree on quality of service parameters like time, costs, reliability, security and trust.

Most of the providers have inflexible pricing, generally limited to flat rates or tariffs based on usage thresholds, and consumers are restricted to offerings from a single provider at a time (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009). The clients are charged according to the quantity of storage they rent in the cloud, the amount of communication that happens between them and the number of server hours they use.

Cloud computing is often compared to the electricity grid as an analogy due to the reach it has to everyone, like the internet, with several service providers. The pricing is similar as well, since users pay per use. Amazon, which is one of the major providers of cloud services, has taken the pricing concept a step further. In December 2009, it launched a new pricing option, where customers bid for its unused computing capacity (Figure 2). Therefore similarly to the electricity, they are beginning to turn computing power into a tradable commodity (The Economist, 2010). Pricing models of those like Amazon make customers think about computing in more economic terms.

There are also several applications that are simply not suitable for the cloud. For example, uploading a high definition video and editing it in the cloud would become a very harsh bargain due to the enormous bandwidth that would take place between the client and the server. Not only the cost, but a bottleneck caused by the latency might be another problem. Applications like for example high frequency trading systems are not suitable for the cloud. The cloud's elasticity and parallelism may be thwarted by data movement costs, the fundamental latency limits of getting into and out of the cloud, or both. Until the cost (and possibly latency) of wide-area data transfer decrease, such applications may be less obvious candidates for the cloud (Armbrust, et al., 2009).

The companies need to balance out the depreciation of their computers and all the costs associated with the systems' maintenance and operations with the costs for cloud computing services and the speed at which they require the results.

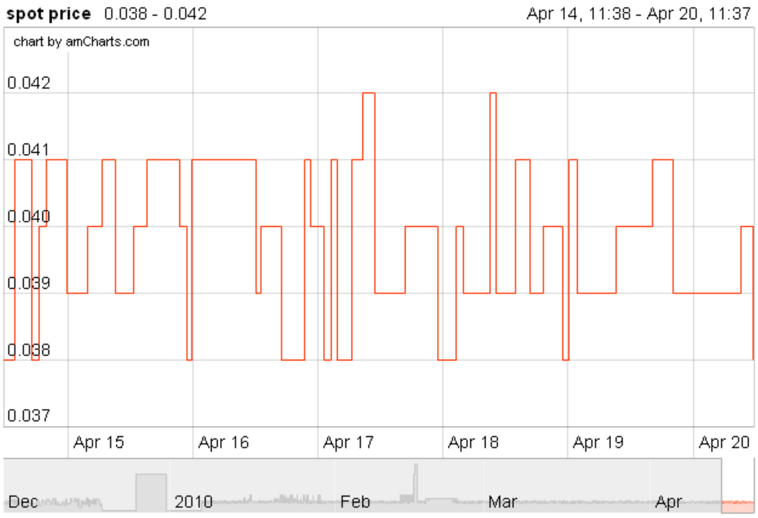


Figure 2: Price of a large virtual Linux machine, $ per hour. Source: cloudexchange.org

2.3.3 Security

In this section we will outline various security risks associated with the use of cloud computing services within companies. Security is usually a combination of confidentiality, integrity and availability. Moreover, the nature of cloud computing makes it necessary to look at three more aspects, and those are compliance, policy and risk (Hobson, 2009).

According to the 2010 State of enterprise security report (Symantec, 2010), cyber security is ranked at the first place as a top risk by 42% of organizations. One of the reasons for this attitude is that 75% of all enterprises which were researched have experienced cyber attacks in the past 12 months (Figure 3).

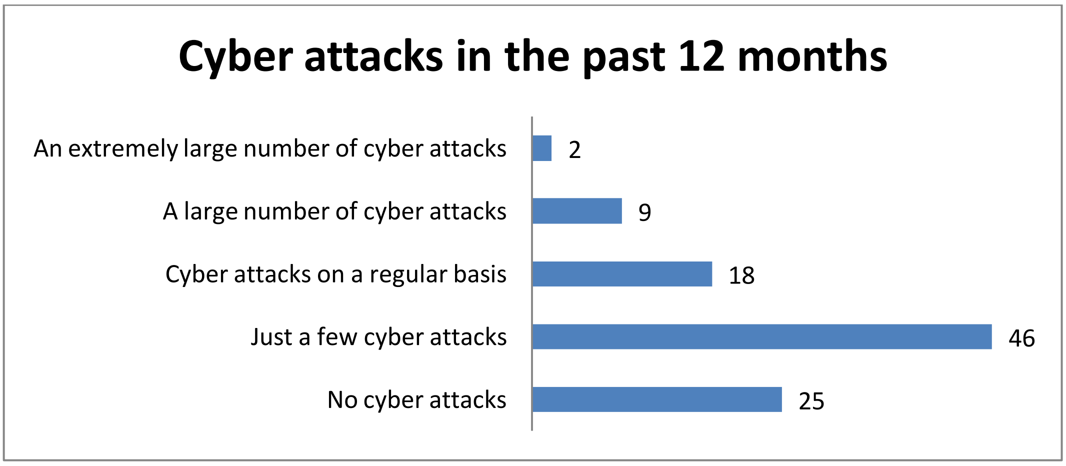


Figure 3: Cyber attacks in the past 12 months. Source: (Symantec, 2010)

Addressing security should be a priority, because the attacks cause serious costs to the enterprise in 92% of the cases. The most common costs are related to lost productivity, lost revenue and lost customer trust.

Distributed systems security involves various aspects of cryptography, secure channels, access control, key management – generation and distribution, authorization, and secure group management (Kshemkalyani & Singhal, 2008).

2.3.3.1 Confidentiality – Integrity – Availability

In cloud computing, the source of threats (Belapurkar, Chakrabarti, Ponnapalli, Varadarajan, Padmanabhuni, & Sundarrajan, 2009) can be divided into three categories: provider, data transfer and clients.

A. Provider, as the host of the service needs to protect the infrastructure from various types of malicious software, but also unauthorized access to confidential data by users and jobs, as well as unauthorized tampering with computational results. The provider faces buffer overflow issues which can be one of the reasons for privilege escalation. Privilege escalation can be horizontal or vertical. In a horizontal escalation, a malicious user tries to get access to another user's private data, or conceal their own identity. In a vertical privilege escalation, the user tries to gain administrator rights to access all the information stored on the host. The applications on the host servers may have to deal with vulnerabilities caused by:
   - Injection attacks: Shell/PHP injection, SQL injection
   - Denial-of-Service attacks
   - DNS attacks
   - Routing attacks
   - Cross-Site scripting
   - Improper session management
   - Improper error handling
   - Improper use of cryptography
   - Insecure configurations issues
   - Denial of service
   - Canonical representation flaws

B. There are many things that can go wrong during the data transfer. The communication between the client and the server can be compromised if the paths are not secure enough. The cloud can be accessed from any place in the world with an internet connection.

Therefore, factors like authentication, authorization and access control, nonrepudiation, data integrity, privacy and trust play a critical role in cloud computing. The communication has to be prepared for various types of attacks:

- Known bug attacks
- XPath and XQuery injection attacks
- Blind XPath attacks
- Cross-site scripting attacks
- WSDL probing
- Enumerating service from WSDL
- Authentication attacks
- Man-in-the-Middle attacks
- SOAP routing attacks
- SOAP attachments virus
- XML signature redirection attacks
- XML attacks
- Schema-based attacks
- UDDI registry attacks

C. The clients need to implement many security measures as well. In the beginning of this chapter we have defined various types of clients that actually access and use the cloud services. Securing the endpoints is critical, because if a malicious user gains control of an endpoint either remotely or physically in case of mobile clients, he might be able to access highly confidential data. From the types of clients, thin ones are the easiest ones to secure. The administrators have full control of what the user can do within the system and if the computer is broken or stolen, it can be almost seamlessly replaced. The most complicated to secure are mobile clients. They add various risks associated with the wireless network, but also the variety of hardware. It is dangerous to allow access to the network to any mobile device, because there might be several unknown bugs that might allow malicious users to overcome the system. There are several security measures which should be executed. The client should use best practices in securing their OS, for example by using cryptography, authentication, avoiding storing passwords, regularly installing security and antivirus updates and so on (Raval & Fichadia, 2007).

2.3.3.2 Compliance

Compliance is now a major business issue. The study of Symantec found that companies are currently exploring 19 separate IT standards or frameworks and are actually using eight of them. The most mentioned are:

- ISO
- HIPAA
- Sarbanes-Oxley
- CIS
- PCI DSS
- ITIL

While considering storing data in the cloud, a company must ask itself several questions like what type of data it is, if it is confidential, or if there are regulations on how and where it can be stored. For example, the PCI DSS standard states regulations regarding the responsibility of storage of credit card data. Appendix A.1.1 of these requirements mandates that no entity other than your organization be able to view your data. Of course, this is desired for all company data stored in the cloud, but one must seek a way of getting reasonable assurance that this will be maintained. It is difficult to trust the cloud provider when the client does not own the infrastructure that holds the data; moreover, it is shared with other clients, possibly even competitors.

So far, there are no regulations for cloud computing service providers. Governments still need to sort out issues of data privacy and ownership of data (Velte, Velte, & Elsenpeter, 2010). The physical location of the datacenter is important from a legal point of view, because local laws affect it. U.S. courts have tended to rule that private data stored in the cloud does not have the same level of protection from law enforcement searches than data stored on a personal computer. Moreover, the U.S. Patriot Act allows the government to have virtually limitless access to any kind of company data. This might not be a problem for data that falls under a regulation, but it is very troubling for intellectual property or proprietary data.

There is hope however, for legal issues related to compliance. The crucial part lies in the negotiation of the SLA with the service provider. During this process, intellectual property rights, privacy laws, liabilities and taxation should be assessed. D. Parrilli writing about legal issues in grid and cloud computing concludes that: *"…legal issues should not be perceived as barriers to invest in Grid and Cloud computing and to start up a successful business. The law, in a very broad sense, does not prevent Grid/Cloud computing from showing all its potential and proving to be innovative technologies able to create new business opportunities, reduce the costs and maximise the profits of* *the users. It is nevertheless true that in some circumstances the legal sources are not fully able to encompass all existing scenarios, included in possible Grid/Cloud-based business opportunities"* (Parrilli, 2010). It was also noted that due to the lack of standardization, in certain cases the users can only rely on trust. This trust is backed up by the provider's reputation, therefore major service providers are recommended. On the other hand, the negotiation process of the SLA with large providers is often unfair for the customers, because they restrict the tailoring of the agreement.

2.3.3.3 Data storage policy

Companies need to find a way to apply their data storage policies to the cloud. A data storage policy is a set of procedures that are implemented to, control and manage data within the organization.

According to Butler Group (Clarke, 2002), an IT analyst company, there are five steps necessary in order to create a correct data storage policy. At first the companies need to establish a data storage budget. If they intend to send some data to the cloud, a cost analysis of this alternative needs to be made. Second step is to assess the data availability requirements. The requirements for availability vary according to the type of data and how often it is being accessed. The third step concerns the measurement of security levels. In this phase the company needs to determine who can have access to a certain kind of data, if it needs to be encrypted and how well. The companies also need to assess legal and governmental requirements, of which some are mentioned above. A more detailed description of this process is discussed in chapter 3.3. Finally, companies need to implement the data policy across the whole organization.

As there are different ways and locations where data is being stored, it is crucial to keep track of it. This becomes difficult in the cloud, because the underlying infrastructure is hidden from the users. Moreover, cloud computing is not the only technology that benefits from a good data storage policy. A way of cataloging the data is needed to be able to search it, which also helps to handle its whole lifecycle. It is also needed for good decision support, email archiving and tiered storage for example.

2.3.3.4 Risk

Risk in this sense is meant as the risk of a provider discontinuing from offering its services; which has already happened in the past. For example in 2008, Amazon's S3 cloud storage service went down two times that year, causing some applications to be offline for even eight hours (Brodkin, 2008). The fact that there are only a few major providers has a large impact on the magnitude such an outage can cause. There are also no standards on the way data is stored in the clouds, making it very difficult to migrate from one to another.

The recent financial crisis has proved us again that no company is too big to be immune from going bankrupt. Clients should keep this in mind and make sure that it is possible to retrieve their data from the cloud and how long this process would take.

2.4 Conclusion

In this chapter we have discussed several aspects of cloud computing. We have shown that although it is a relatively new business model, the idea itself is not completely new. Being just another step in the evolution of distributed systems, it was waiting for the IT industry to mature enough, making the concept realizable. At this moment, cloud computing is regarded to be in its infancy making solutions like Software plus Service to be more appealing than simple SaaS services. Although consumers enjoy using these tools, for professionals it is still more convenient to use standalone applications which offer better functionality. Combining them with cloud computing only makes them better.

From an economic perspective, we can say that cloud computing promises to use IT more efficiently, making it very attractive for companies which would like to cut costs on their private IS. On the other hand, not only can they cut costs, but they can become more effective. Thanks to the elasticity of the service, they are able to respond to market changes more quickly, giving them a competitive advantage. Cloud computing can increase competitiveness, as it also gives a chance to start-up companies. They can begin operating with relatively small capital investments in an IS and scale it according to their needs as they grow.

Looking more closely at the concept of cloud computing, we realize that it is not so much of a green field as it may seem at first. Most of the problems associated with the communication path between the client and the provider have been dealt with in the past, though there are several new problems that still do not have clear solutions. These are mainly security risks associated with compliance and policy. In this chapter we have discussed these risks and we have laid down some current issues with the legislation. It is clear that until these issues have been sorted out, companies cannot allow all of their data to migrate into the cloud. Moreover, even though these legal issues regarding the ownership of data in the cloud would be sorted out, companies will still avoid storing their most confidential or proprietary data here due to the fear of leakage.

The risks associated with cloud computing services are so diverse, that it will probably take some time before a company could replace its entire IS with cloud computing. On the other hand, the benefits are clear and have great potential, making the companies want to experiment with it at least partially. Defining the value of data is absolutely necessary in the process of preventing unwanted data to be exposed in the cloud. The State of enterprise security research done by Symantec which has been quoted in this chapter recommends that companies should take a content-aware approach to protecting information as it is the key to knowing where the information is, who has access to it and how it flows within and outside the company. It also recommends companies to develop IT policies that would span across the whole organization and enforce them through built-in automation and workflow.

A data value management would solve the currently unresolved issues with compliance and similar problems, and it would still allow the companies to enjoy the benefits of for example better team collaboration, disaster recovery planning and expenses cuts. In the next chapter, we will look at data classification as a tool for data value management.

3 : Data Value Management

3.1 Introduction

Moving on from cloud computing, we now concentrate on data value management. In this chapter, we present a literature review of the theoretical background behind the task of data classification. We describe the processes required as a preparation for data classification and discuss various technical solutions together with their drawbacks.

3.2 Motivation

Humans tend to give a label to everything. In music for example, we can distinguish between separate genres because members of these groups carry certain similarities. We can go further and label music according to the author, year of publishing, beats per minute, and so on. This kind of classification helps us to bring order to chaos. Without classification it would be difficult to understand connections between things and our knowledge would be quite ineffective. *Data classification is fundamental to asset management, risk assessment and the strategic use of security controls within the IT infrastructure of any organization* (Etges, CISA, CISSP, & McNeil, 2006).

The vision of data classification based on business value was not implemented successfully in the past, because it was difficult to automate this process; thus required a lot of manual intervention. Instead, the classification was mainly age-based where the date of creation and modification were the only criterions as this was much easier to automate (Vellante & Floyer, 2008). As we have mentioned in the previous chapter, this kind of approach is naturally insufficient for information lifecycle management (ILM), tiered storage, email archiving, decision support and most of all the new compliance issues related to cloud computing storage.

The bullish economy in the 90s caused the companies to solve their growing storage capacity needs by simply adding more hardware to their systems. Data management mostly dealt with structured data like databases, which accounts for only a fraction of data compared to unstructured data[5] created as an output of the users (Toigo, 2005).

There are some obvious motivations for improving data classification models in information systems. First of all, storage capacity requirements grow with unmanaged data, so costs rise as well. It is important to place the right data on the right platform, at the right time. The platform needs to be assessed from a performance, accessibility and cost perspective. This will enable the delivery of the

[5] The exact numbers differ, but there is a general acceptance of a ratio of approximately 20:80 percent (Grimes, 2008).

best service to the organization at the lowest possible price. Jon Toigo (2005) from Data Management Institute summarizes the benefits of data classification into three perspectives:

- Cost-savings; can be achieved by reduced total cost of ownership of storage and by managing data that is no longer of any use to the organization.
- Risk Reduction; can be achieved by reducing the downtime and data inaccessibility through proactive problem resolution. This can be done by segregating critical data and shrinking backup volumes.
- Business Process Improvement; can be achieved by creating an effective data model of IT support costs based upon business processes, applications and workflows where the data produced and used by each application is discovered. Such a model can be of enormous probative and predictive value to business decision makers. It has also audit implications, where *classification can enable the reconstruction of a continuum of organizational activities performed and decisions made over a period of time* (Vellante, 2007).

3.3 Data classification strategy

We have introduced the need for applying a data storage policy within an IS in the previous chapter. Before we can determine how we are going to categorize data, first we need to determine what we are actually looking for. Therefore, all lines of business should cooperate with IT and perform several tasks. These steps are critical in the process of the data classification strategy, because businesses have to understand what kind of data they produce (Figure 4).

- First of all, all business processes should be identified and deconstructed into their tasks and workflows.
- Secondly, applications that support these workflows should be identified and their data flows mapped.
- Finally, during the process of business workflow declassification, we should begin to see some commonalities between the data and thus several classes should be established. Based on the findings, critical business information should also be identified. All the data should be analyzed to determine its management requirements and enable to discern common schemes of classification (Toigo, 2005).

Once we know the classes of data which are occurring within a business IS, the next step is to express security characteristics such as ownership, liability and control of data. Although IT is usually responsible for maintaining different business applications, business information does not belong to IT (Appleyard, 2003). The Control Objectives for Information and related Technology (COBIT) framework[6] recommends that management should implement procedures that would classify all the data in terms of security, whereas the data owner should decide about this. Data classification is a part of ILM which adopted some classification levels from military systems. Categories of security like for example private, sensitive, critical and confidential name various levels where different rules apply. The names given to these levels are not important, as long as data sets that apply meaning to business operations are established. The classification levels should be based on reasons why the data are important to business, although IT does not have this knowledge in most cases (Etges, CISA, CISSP, & McNeil, 2006). For this reason the communication between the IT and data owners is critical in this phase of the classification strategy. The model has to be developed in each company individually to satisfy their needs and can hardly be supplied as an off-the-shelf solution. It might involve conducting interviews and surveys with relevant people responsible for each business process.

After the model of data classification is completed, a data storage policy as mentioned in the previous chapter (subsection 2.3.3.3) can be derived from it. This is then used to classify data about to be sent into the cloud. It is also vital to communicate the data storage policy throughout the organization so that the employees understand what kinds of data they are working with, and what they are allowed to do with them.

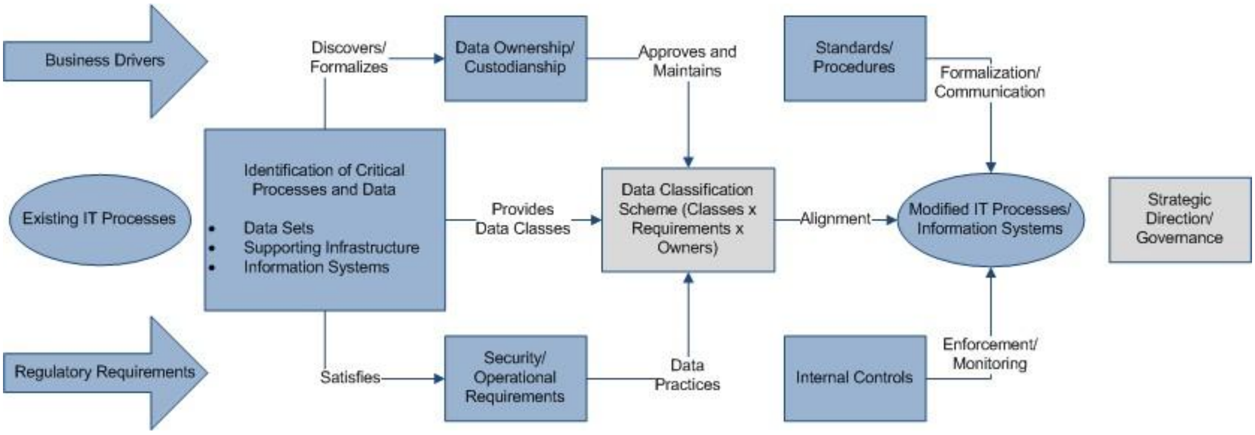

Business Drivers
Discovers/ Formalizes
Data Ownership/ Custodianship
Approves and Maintains
Standards/ Procedures
Formalization/ Communication
Existing IT Processes
Identification of Critical Processes and Data
Data Sets
Supporting Infrastructure
Information Systems
Provides Data Classes
Data Classification Scheme (Classes x Requirements x Owners)
Alignment
Modified IT Processes/ Information Systems
Strategic Direction/ Governance
Regulatory Requirements
Satisfies
Security/ Operational Requirements
Data Practices
Internal Controls
Enforcement/ Monitoring

Figure 4: Overview of the data classification process. Source: (Etges, CISA, CISSP, & McNeil, 2006)

3.4 Business and security criteria for data classification

There are several criteria which need to be considered when deciding about the data classification model. Similarly like in chapter 2 about cloud computing, we asses these requirements from security aspects like confidentiality, integrity and availability, but since we are now describing data and not an IS, additional aspects like auditability, access and authentication apply. The following subsections are based on the paper by Edges et al. (Understanding Data Classification Based on Business and

[6] Here we refer to COBIT control objectives in DS5: Ensure systems security

Security Requirements, 2006), although these have been adjusted with some cloud computing examples.

3.4.1 Access and authentication

Not everyone really needs access to all information. The access can be further distinguished by people who need access to information for regular business operations, who need it only for support and maintenance and who need access because of audit operations. Collectively, these levels are referred to as access rights. The goal of authentication is to give reasonable assurance that the user is really who he claims to be. A certain level of trust has to be established in order for the user to enjoy his access rights.

The COBIT framework recommends that the owner of the information who is also accountable for anything that happens to the data should be supported by a formal approval and authorization mechanism. This is of course unless unlimited access to certain data is allowed.

3.4.2 Confidentiality

This is one of the most important aspects of the classification scheme, especially when considering integrating cloud computing services within a company IS. Once sensitive information is identified, it is important to determine where it will exist. This includes questions about storage, transmission and manipulation of data. These questions should be answered according to legal and compliance requirements which are mentioned in chapter 2. A company might want to determine where their proprietary information can be archived and so on.

3.4.3 Privacy

Similarly to confidentiality, there are some legal and compliance requirements for private data. There are rules on what a user is allowed to do with certain private data. Controls for this ability include identification and authentication of the user, access controls, integrity controls to make sure the changes are within limits, and logging for auditing of the changes made.

3.4.4 Availability

Depending on how critical certain data is, availability requirements vary. This affects decisions on how often it should be backed-up, where, and if it should be made redundant. From an ILM perspective, it should be also determined when to destroy the data. Decision should be made according to the acceptable downtime but should be also based on the fact how often the data is accessed. For example, archiving a very large file in the cloud would not be cost effective, if it would require very frequent access. This is due to the pricing scheme, but it depends greatly on the SLA between the provider and the company. On the other hand, storing large data sets which are legally required to be archived for several years due to audit reasons might be very beneficial as long as other requirements are satisfied.

3.4.5 Ownership and distribution

The controls of confidentiality, integrity and availability might not protect copyrighted or proprietary data sufficiently from unauthorized copy and distribution. Some companies therefore implement watermarking or cryptographic techniques to protect themselves from piracy. Distribution channels for such data should also be established to minimize the risk of data leakage. A company might implement rules for example on the distribution of certain files through emails or instant messengers.

3.4.6 Integrity

Integrity ensures that data is protected from unauthorized changes. This can be ensured through several techniques, for example by creating hashes of stored data and comparing them with newer hashes of the same files. Secured communication should be ensured during transit by for example secured Virtual Private Networks. Data should be checked for accuracy during manipulation to prevent fraud. There should also be rules on using “lossy” compression techniques on files that are not text based.

3.4.7 Data retention

Compliance frameworks require companies to archive files relevant for the audit for several years. Files that fall into this category must be identified to protect them from deletion. This includes personal records, emails and financial reports. Files necessary for getting access to these files like for example encryption keys, credentials but also versions of software under which files were created need to be archived as well.

3.4.8 Auditability

It must be identifiable who created, viewed and modified data. If this is implemented, it is possible to trace back everything that has been done with a file throughout its entire lifecycle. The level of detail of these logs can vary but generally they are required for auditing and control.

3.5 Existing solutions

In his paper *Data Classification in Distributed Computing Environments* (Toigo, 2005), Toigo claimed that at the time, no vendor had a comprehensive data value management solution, which would address all the criteria as described in section 3.4 and satisfy all the needs of a true ILM. He claimed that Enterprise Content Management systems do not as a rule capture the files created by knowledge workers in their workflow-based data movers, i.e. they do not classify data in relation to profiles of users who have created it. Database management systems are insufficient because structured data accounts for only a small fraction of data within an organization, and for the same reason Hierarchical Storage Management, email management, archiving and backup are not complete solutions.

Although the solutions to address the problems of data classification have been in short supply, today the situation has somewhat improved (Wendt, 2010). The IT security market is beginning to offer a solution which uses specific techniques to monitor and identify confidential information, which is called Data Loss Prevention (DLP). Mr. Wendt, an IT analyst, claims that DLP could be used for example to scan through large archives in the cloud, in order to locate and delete confidential information which is not needed anymore.

All major security vendors offer some kind of a DLP solution, including Symantec, McAfee, Cisco, Vericept, Fidelis and many others. As this market has not yet fully matured, similarly to cloud computing, there seems to be some confusion around the definition and the exact term to call it. As Tony Zirnoon from McAfee writes (Zirnoon, 2009), some security vendors call this system ILP (Information Leakage Prevention), ILDP (Information Leakage Detection and Prevention), CMF (Content Monitoring and Filtering), and even EPS (Extrusion Prevention System) as an analogy to Intrusion Prevention System. He further proposes the use of the following definition of DLP: *[DLP systems are] products that, based on central policies, identify, monitor, and protect data at rest, in motion, and in use, through deep content analysis.*

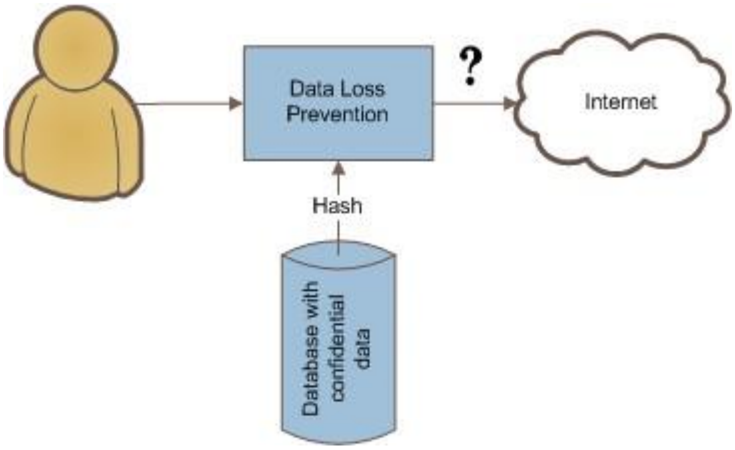


Figure 5: Data Loss Prevention Scheme. Source: (Leffel, 2010)

The principle of how a DLP system works (Leffel, 2010), (Figure 5) is quite simple. First it creates hashes of all the data that has some special value to the company, a process referred to as fingerprinting. This is done to avoid any actual copying of confidential data. When a user creates some content and wants to send it out of the company IS, it is scanned by the DLP system out of which hashes are created and compared to the hashes that have already been prepared. If any of its content matches, some preventive action takes place. Because of this technique, the system is able to identify confidential information in many forms, no matter if it is in a text document or a spreadsheet. Once a conflict is detected, it might warn the user that the file he is about to send out contains confidential data and it might quarantine it until it is approved. The DLP system can not only monitor data that moves out of the IS, but it can also analyze files that move within the IS and see where they are stored. It can check if a file is allowed to be stored on a laptop for example, or if a user can print it. The system can implement various controls based on the requirements of the data classification model as described in the previous section (3.4). However, we can make several observations of limitations for the use as a data classification solution for cloud computing use.

First of all, it is obvious from the scheme (Figure 5) that in order for the DLP system to work, it has to have a clearly defined set of data (customer data, corporate data or intellectual property) which it should look out for. For example, a database containing customer data with credit card numbers would cause the system to look for strings with the same length and content as in the database, but the system would not know that other 14 digit numbers could also be credit card numbers. Therefore the system lacks the ability to *learn* and *predict*, thus requires a lot of administration and tweaking. If it would have had these capabilities, it could bring additional indirect benefits like for example *anomaly detection*.

The DLP system can only work with textual files and data, but confidential data does not have to be only in these forms. Files that were not "*digitally born*" like for example scanned documents, audio files, photos or PDF files would escape the data classification scheme.

These types of files occur in all sorts of businesses. An unstructured data analysis report of a finance midmarket client (Williams, 2010) has shown that although 50% of the files created by users are office documents, 10% are PDF files, around 6% are still pictures and there are still 16% of other file types (Figure 6).

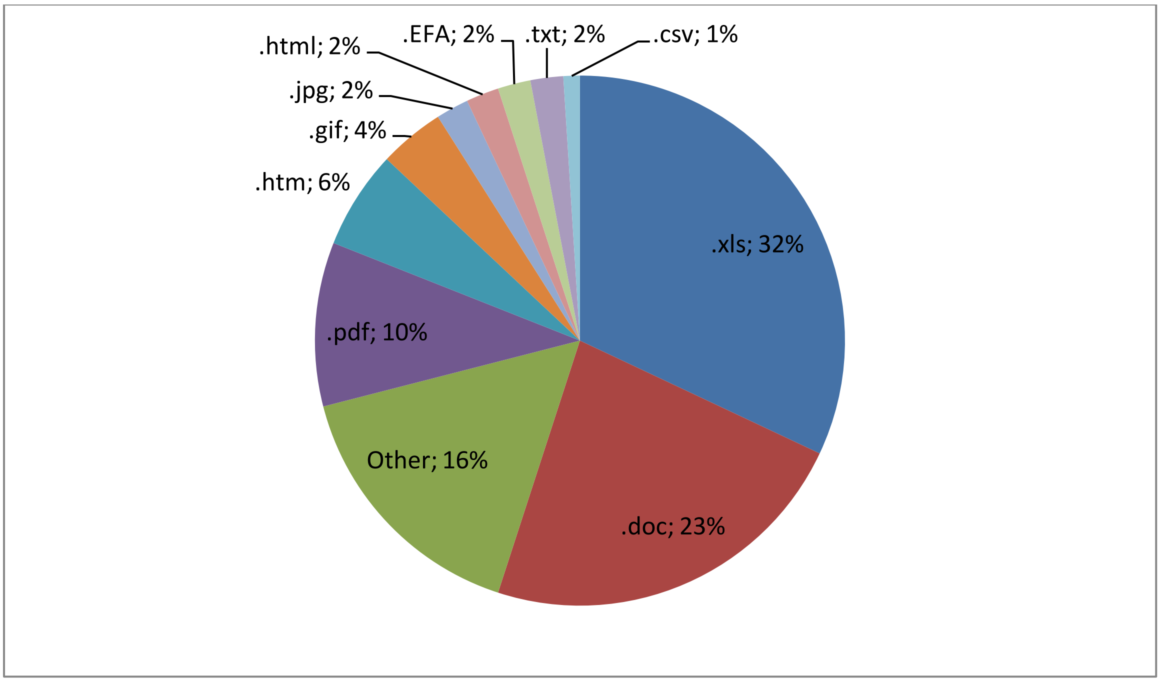


Figure 6: Top ten files by count, showing where the majority of user effort and file creation occurs. Source: (Williams, 2010)

The task of the DLP systems is very complex already, but in order for the system to be more thorough in the classification, it needs to use more complex techniques for deep content analysis. The DLP system focuses on identifying files with confidential information where they are found the most often and the ways they travel out of the company, for example in the form of email attachments. Such a solution with a focus on data traveling out of the company might not give enough reasonable assurance for security of data traveling within the company IS and deciding if it can be used in the cloud.

3.6 Conclusion

In the past, businesses did not focus enough on various classification techniques when developing their IS. It is much more easier to implement data classification proactively than retrospectively to an already well functioning system that became very complex with time. Unfortunately, due to the previous neglect, most of the systems need some kind of improvement of their insufficient classification procedures. Such a task can be endless and daunting if done manually; therefore a great degree of automation is necessary which would also be useful for new data due to the amounts that are being constantly created. There are calls for automation of the compliance requirements and capabilities to anticipate threats before they happen (Symantec, 2010), and also automatic information management (Hey & Trefethen, 2003). Some manual intervention is still necessary though, because relying solely on machines might be dangerous. These automatic classification techniques should therefore complement a user-driven approach.

It has been indicated in this chapter that there is no complete solution which would be able to reliably classify all the data that is being created and moves within an organization. A data storage policy might be more complex than the technical solutions implementing it. The closest to satisfying the requirements for classifying data due to cloud computing integration, are probably DLP systems. However, they were not invented for this purposes therefore their concentration is more on how data leaves the organization instead of how it moves within. Data which is being moved to the cloud does not have to be considered as leaking, but it should be tested on compliance, regulatory and internal policy requirements. There is a need for a classification system that would encompass more complex deep content analysis techniques and also become more autonomous in the decision process. Moreover, a variety of solutions should be used in combination, to be able to deal with all threats (Takebayashi, Tsuda, Hasebe, & Masuoka, 2010).

While in this chapter we have been mostly talking about the theory of data value management and which factors a data classification system should take into account, the next chapter will talk about the components such a system should have in order to satisfy the data storage policy requirements.

4 : Automatic Data Classification

4.1 Introduction

Chapter 3 has concluded with the message that there is currently a lack of a complete data classification system which would fully satisfy the requirements for cloud computing integration and use. Current data loss protection systems can detect sensitive information only if they were previously told what to look for. This is a major drawback that requires constant updating due to the ever increasing amounts of data. Unless the database is updated, it might result in a high number of false-negatives, i.e. failing to detect sensitive information. Moreover, such a system might find it difficult to accept that a certain user might be allowed to send files containing sensitive information around the network, resulting in a high number of false-positives, i.e. classifying sensitive files too strictly. High number of these errors can negatively affect the user-friendliness of the system, ultimately causing even more damage when users stop trusting it and begin ignoring the warnings.

The only way to solve these problems is to use more comprehensive algorithms with learning capabilities, combine technology and create greater independence from user interference. The goal of this chapter is to present information on what an automatic data classification system should be composed of. The intention is to have a system that could be integrated into a company IS with the goal to use cloud computing services seamlessly and safely and overcome the drawbacks of the DLP systems described in chapter 3.

4.2 Metadata

Reading through the previous chapters, it should be clear that the goal of data classification is to provide better knowledge of stored data. This kind of knowledge is generally referred to as metadata. A simple and widely accepted definition says that metadata is data about data (Masinter, Archer, Connolly, Hammer-Lahav, & Denenberg, 2001), (Hey & Trefethen, 2003). On the other hand, there are opinions that this definition is outdated and propose a more sophisticated one (Execution MiH, 2010):

Metadata is 'data, information and knowledge' on the 'data, information and knowledge' of an organization, both from physical and logical perspective; whereas the difference between data and information is, that information adds some kind of semantic meaning to it. Data is transformed into information through analytical abilities, technical skills, experience and intuition, and ultimately enhances decision-making (Barsky & Catanach, 2010). The knowledge referred to in the above definition is the integration of data with business processes and workflows as described in chapter

3.3. All in all, it leads to knowledge on how to handle specific data. Therefore, metadata should be able to tell us several important aspects about flowing and even newly created data:

- Data type
- Location of data
- Data lifecycle audit information
- Implemented access levels
- Implemented protection levels
- Compliance regulations requirements

There are also some other (technical) requirements for metadata (Nilsson, Palmér, & Naeve, 2002):

- Metadata is objective data about data.
- Metadata for a resource is produced only once.
- Metadata must have a logically defined semantics.
- Metadata can be described by metadata documents.
- Metadata must be the digital version of library indexing systems.
- Metadata is machine-readable data about data.

To satisfy the requirement that metadata must be machine-readable and thus improve searching through indexing, it has to be properly encoded in a meta-language. The most common solution to this problem is the XML (Extensible Markup Language) which uses tags to describe everything. A more thorough technical description of XML and relevant protocols for metadata recording is out of the scope of this thesis.

It is logical that the more metadata to every file, the more data is created in general. However if used correctly, it can deliver benefits which were described in chapter 3 (section 3.2), which include considerable increase in storage efficiency by e.g. removing redundant data and by archiving some of it in the cloud, ultimately leading to lower costs.

4.3 DAR

Once we have established that we are going to represent the classification information through metadata, we now look how it can be generated. Since most of the data is unstructured, this is the focus of our research and of the science of Document Analysis and Recognition (DAR).

DAR aims at the automatic extraction of information presented on paper and initially addressed to human comprehension. The desired output of DAR systems is usually in a suitable symbolic representation that can subsequently be processed by computers (Marinai, 2008). The fact that DAR is able to understand documents that were not digitally born can be used as an extension to existing DLP systems to widen the scope of analyzed files. The techniques of semantic information extraction are better than simple search for certain keywords in a document. This is because with semantics, the computer is able to distinguish between a formal and an informal letter for example, which should lead to enhanced decision making.

The process of DAR comprises of three steps (Marinai, 2008):

- *Pre-processing*. In case of scanned documents or photos, this processing takes care of improving the quality of the image. It might involve clearing noise or anomalies that might affect the result. In case of documents that have digital roots, this process can be omitted because they do not have these flaws.
- *Object segmentation*. This phase is a layout analysis where independent regions in the document are identified. Text of an email for example, consists of a region where the salutation is, several paragraphs, a closing and a region where the signature is.
- *Object recognition*. This phase recognizes the type of object which was identified in the previous phase.
- *Post-processing*. This phase checks the result of classification based on contextual information.

The DOMINUS (Esposito, Ferilli, Basile, & Di Mauro, 2008) system is a framework that exploits intelligent techniques to support different tasks of automatic document processing from acquisition to indexing and from categorization to storing and retrieval. It was designed to be embedded as a document management engine into various digital library systems.

The system uses the DAR techniques described above to extract metadata information from scanned documents and PDF files. It is fully incremental, which means that it is able to modify the existing model and learning can also start from an empty theory. The learning system of the DOMINUS features a closed loop architecture, which means that evaluation of performance is used to activate the theory revision phase. The system was demonstrated on classification of scientific journals and the classification itself is performed by matching the layout structure of their first page against automatically learned models of classes of documents. Concrete training or evaluation algorithms are not described in detail, but with the right methods, this model could be probably extended to analyze almost any kind of document with text in it. Moreover, if the system would analyze the content of all pages of a document, it would be able to semantically determine how to classify it in terms of security.

The automation of the process where we gather information about data is possible through various algorithms. Depending on the algorithm, the model, or the data, we might call such an activity pattern classification, statistical pattern recognition, information retrieval, machine learning, data mining, or statistical learning (Maloof, 2006). Even though there are differences between these concepts, we will describe their similarities to better understand the overall concept of machine learning or data mining.

4.4 Machine learning and data mining

Here we present several algorithms common for machine learning and data mining which are used in the object recognition and post-processing phase of DAR. These techniques can be applied at various levels of the classification process to improve the results. For example, in case of DAR, the goal is to create metadata that would sufficiently and reliably describe a certain file. During the object recognition phase, machine learning can be used to understand if the document is a formal letter, a contract, or a blueprint for example. Once the metadata is created, we can apply some learning and mining techniques again, to determine what can be done with the file, i.e. to classify it.

4.4.1 Examples

In order for data mining and machine learning to work, the algorithms need to analyze a certain set of source data. Raw data is rarely suitable for learning or mining, therefore we need to create sets of examples which these algorithms should analyze. An example consists of a set of attributes and their values whereas the operation of raw data transformation into examples includes many operations including the following, which are referenced in (Maloof, 2006).

- Adding a new attribute calculated from others
- Mapping values of numeric attributes to the range
- Mapping values of numeric attributes to discrete values
- Predicting missing attribute values based on other values
- Removing attributes irrelevant for prediction
- Selecting examples relevant for prediction
- Relabeling mislabeled examples

The process of object recognition involves the understanding of text regions that were identified in a file. An example may consist of attributes that would say if the region contains specific words or phrases like for example “Dear” or “Hi” in case of a salutation whereas the location of the region with reference to the whole page might be considered as well. Using the examples, algorithms are applied to build desired models. The desired model in our case is a classification that would address compliance, regulatory, policy and other issues.

4.4.2 Algorithms

Learning and mining algorithms have three components:

- Representation
- Learning element
- Performance element

The representation is a formalism which determines which models can be built. The learning element builds a model from a set of examples and the performance model applies the model to new observations that need to be classified.

The model is a compact summary of the examples usually and it is often easier to analyze the model than the examples themselves. The model also often generalizes the examples, which is a desirable property, because we do not need to describe all possible examples to build one.

The following examples are algorithms that are commonly used in data mining and machine learning as listed in the book by Maloof (Machine Learning and Data Mining for Computer Security, 2006). A system architect building an automatic data classification model might also want to look into other (non-discussed) algorithms (e.g. neural networks, support vector machines, etc.).

4.4.2.1 *Instance-Based Learning (Aha, Kibler, & Albert, 1991)*

Learning consists of storing examples as the learner's concept description. When classifying an observation, the performance element computes the distance between the observation and every stored example. For numerical attributes, Euclidean distance can be calculated and tally mismatches can be used for symbolic attributes. There are variations where the *k* closest instances (IB*k*) or *k* nearest neighbors (*k*-NN) are found and the class with the majority vote is selected as the final decision.

4.4.2.2 *Naïve Bayes (Domingos & Pazzani, 1997)*

In this case the learner's concept description is the prior probability of each class and the conditional probability of each attribute value given the class. By counting the frequencies of occurrence, the learning element estimates these probabilities. The decision is given according to the class with highest probability.

4.4.2.3 *Kernel Density Estimation (John & Langley, 1995)*

This estimator is similar to Naïve Bayes, but in addition to storing the prior probabilities, it also stores each example. The performance element estimates the probability of a value given the class by averaging over Gaussian kernels centered on each stored example.

4.4.2.4 *Learning Coefficients of a Linear Function (Minsky & Papert, 1988)*

This algorithm uses a linear classifier which is restricted for binary input or two-class problems. It has been proved that if the positive and negative examples are *linearly separable*, the algorithm is guaranteed to converge to a solution after a sufficient number of iterations. There are some circumstances though under which the algorithm can only find an approximate solution.

4.4.2.5 *Learning Decision Rules (Holte, 1993)*

The decision rule consists of an antecedent and a consequent. In other words, if an observation satisfies a list of tests, a class label is given to it. It can be modified in such a way that if an observation does not match any rule, the decision is generalized to the class with the majority of satisfied conditions.

4.4.2.6 *Learning Decision Trees (Quinlan, 1993)*

The principle of the learning element of this algorithm is to create a tree recursively by selecting the attribute that best splits the examples into their proper classes. It then creates child nodes for each value of the selected attribute, and distributes the examples to these child nodes based on the values of the selected attribute. After that the algorithm removes the selected attribute from further consideration and repeats for each child node until it produces nodes containing examples of the same class.

4.4.2.7 *Mining Association Rules*

This algorithm was motivated by analyses where patterns in baskets of purchased items had to be discovered. This algorithm produces an exhaustive set of all rules that satisfy a consequent similarly to *Learning Decision Rules*. Moreover, the Apriori algorithm (Agrawal & Srikant, 1994) can generate all association rules that satisfy three constrains: the number of attribute values to be used to form the rule, the level of support, and the level of confidence the rule must have.

4.4.3 Evaluation

After a model is built using a learning or mining algorithm, it is required to determine how well the induced model approximates the true model. Since the model is only an assumption, there is some uncertainty that needs to be tested for quality.

There are several evaluation schemes that test the models including: hold-out, leave-one-out, cross-validation and bootstrap. These methods vary, but they usually involve dividing the data set into one or more training and testing sets. An algorithm is applied on the training set and results are evaluated on the testing set to compute a performance metric.

There are several simple measures of performance including:

- Accuracy expressed as percentage
- Error rate
- True-positive rate
- True-negative rate
- False-negative rate
- False-positive rate

For an evaluation of an algorithm over a range of possible operating scenarios, a receiver operating characteristic (ROC) analysis is used. It is a plot of a model's true-positive rate against its false-positive rate (Figure 7), whereas the area under the curve can be regarded as a single measure of performance. A larger area represents better performance.

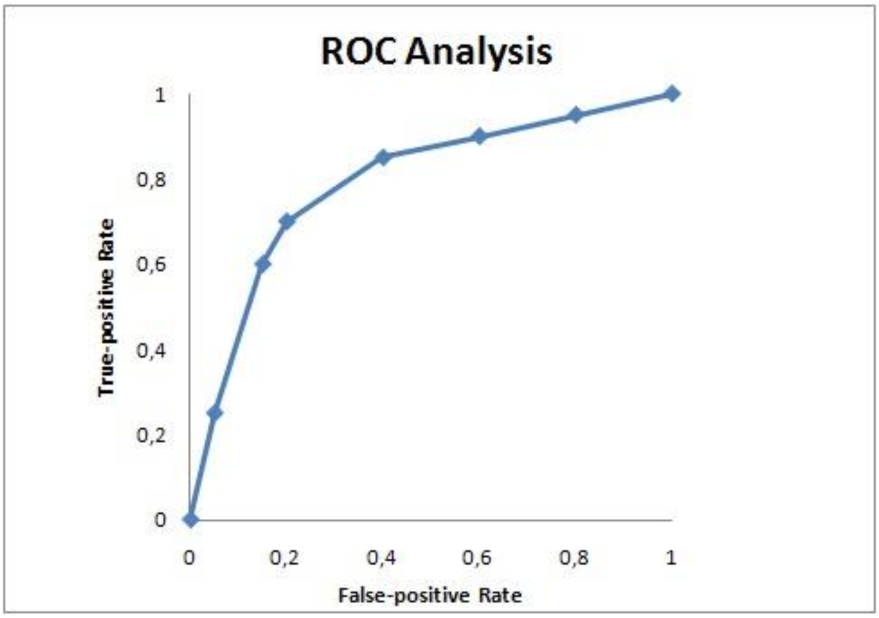


Figure 7: Example of a ROC analysis. Source: (Maloof, 2006)

4.5 Multimedia files

Audiovisual files can also contain confidential or sensitive information. A business meeting could be recorded on an audio file and archived for future reference for example. The fact that multimedia files do not natively contain any text makes it impossible to simply search for specific keywords which could give a clue what the file contains. For this reason a multimedia content description standard has been standardized by the International Organization for Standardization (ISO) now known as MPEG-7 (ISO, 2002).

This standard does not deal with the actual encoding of audiovisual files, but serves as a complementary tool to existing MPEG standards. This makes it more universal and can be attached to various types of files even including still pictures, graphics, 3D models, audio, speech, video, and combinations of these. The objective of the metadata created by the standard is to enable efficient searching, filtering and content identification (Hunter, 1999). MPEG-7 allows descriptions of audio–visual content at different perceptual and semantic levels. It is expected that low-level features (such as color and structural features) can be automatically extracted in fully automatic ways, whereas high-level features—in particular, those that describe semantic information—need (much) more human interaction or textual annotation by humans or agents (Chang, Sikora, & Puri, 2001). In practice this means that in order to be able to match certain features in the analyzed file, an original file (example) is necessary for comparison.

The strength of the MPEG-7 lies in its efficiency and speed. In case of a video file for example, it works by extracting a hash from the video by comparing the brightness of up to 360 areas in a frame whereas each area is assigned a value -1, 0, or 1. This method is very efficient because it identifies content even if the original video has been altered.

Evaluating the frames in a simple way makes the creation and analysis of metadata very fast. An ordinary PC can be used with the system to scour through 1,000 hours of video in a second (Economist, 2010). For example YouTube, a video sharing website, uses this or similar technology to discover unauthorized uploads of copyrighted material (YouTube, 2010). The Google search engine on the other hand uses image description algorithms to search for similar images using pictures and not words (Google, 2009).

Another benefit of MPEG-7 is that the actual descriptors do not need to be attached physically to the actual file they describe and they can be created additionally without the need to change the format of existing ones.

4.6 Automatic data classification system proposition

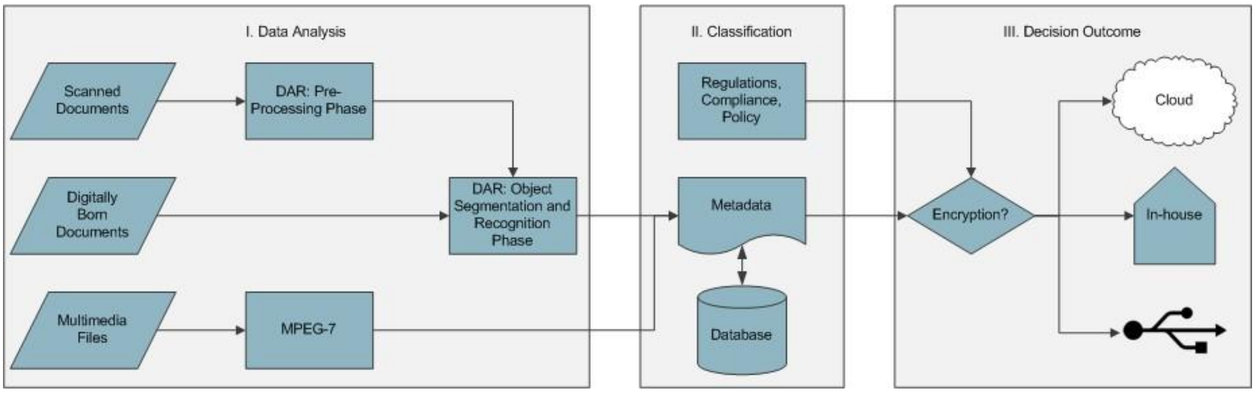


Figure 8: Automatic Data Classification System Proposition. Source: Own compilation.

I. The proposed automatic data classification system (Figure 8) combines techniques of DAR, DLP systems and adds functionality for multimedia files. In the first phase of data analysis, the system has to decide which techniques it will apply based on the source file. In case of textual files it would attempt to extract metadata from them by using DAR methods with the complementary use of data mining and machine learning techniques. In case the file is digitally born, which in practice accounts for most of the files, the model would directly apply the Object Segmentation and later on the Object Recognition phase. If the file is a scanned document, or if the text is converted to a form in which the system is not directly able to copy it, the Pre-processing phase has to prepare it to recognize the text the file contains. The system should also be able to open archives in zip or similar format.

Non-textual files in a form of audio, video, still picture formats and even graphics can also be analyzed. These types are described using the MPEG-7 format which also creates metadata which can be further processed in the classification phase.

II. Once the metadata has been created, this information can be used to classify the files. Inspired from the DLP system architecture, a database containing hashes of specific bits of confidential information would be used to bring additional functionality to the classification. It would also contain MPEG-7 files of elements that need to be discovered in the analyzed multimedia files. In case of textual files, an analysis same as in the DLP systems would take place independently from the DAR analysis to enhance the classification decision making.

III. In the final phase, the proposed model should already know what type of file it is dealing with, and also if it contains confidential information that needs to be protected. Again, in order to automate the whole process of classification, data mining and machine learning techniques are applied to build decision models. The goal of the models is to take into account the data classification strategy based upon regulations, compliance and policy issues as discussed in chapter 3 and decide if the file needs to be encrypted. After that the model decides about the allowed physical locations. As a result a decision would take place to determine if the file can be sent to the cloud, if it has to stay in-house or even if it can be stored on a removable disc, be printed and so on. If the decision model does not have sufficient assurance, or if it cannot make the decision for some other reason, it would require human interaction of the user to decide and classify the file himself. Because of the fact that the decision model has incremental functionalities, it would extend and improve automatically, when users add more hashes of sensitive data into the database, or by simply gaining more experience with the classification.

It is difficult to determine the speed of such a classification system, because some algorithms are faster than others, therefore testing would have to be conducted to find an optimal balance between speed and efficiency. In any case, such system would have to be placed on a server within the company's IS, i.e. in-house. It would have to monitor all network traffic in a similar fashion to DLP systems. Because of the system's learning capabilities, it should not only serve as a gateway to cloud storage or as a data loss protection solution, but it would also have potential to be tweaked to detect anomalies within the IS and serve as an early warning system.

Implementing such a system to a well established IS is a complex task that would involve the participation of departments all across the company. First of all a data classification strategy would need to be followed so that specific classification rules can be determined. Models of classification should be well considered so that the system could learn and create its own decision models from solid examples. Such an implementation would probably require a relatively long time and a lot of supervision in the early stages of development. However, if the details are well thought through, this proposed combined data classification system should pay off by providing results like no other current classification solution can on its own.

4.7 Conclusion

Building a solid data classification model requires the combination of various systems which have to cooperate in the heading towards a common goal. The system has to be integrated into the IS seamlessly, therefore requires as little interaction from the users as possible. This independence can only be achieved if the system has incremental capabilities, because it is not possible to determine in advance all the possible examples of classes.

At the moment we are unaware of a classification system that would be complete and able to deal with all sorts of data types. For this reason, the proposed system should serve as a descriptive study for a system architect in charge of implementing such a solution. The important positive note is that all the technology is already available today; it only needs to be combined correctly.

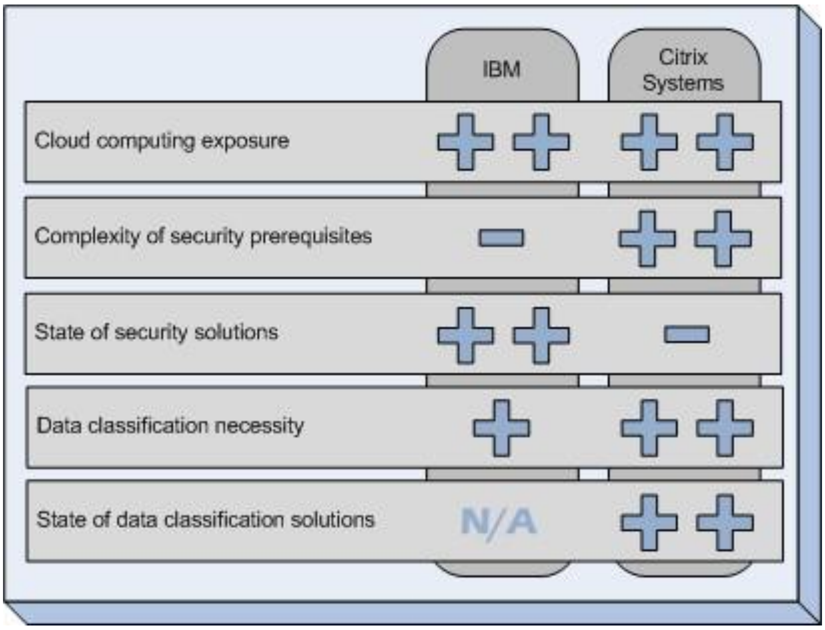
5 : Empirical study

5.1 Introduction

The goal of this chapter is to verify the collected knowledge with people who have real life experience with cloud computing. Two ways of gathering data took place. In one type, open questions were given to respondents in the form of an interview. This was done intentionally, because we wanted to see if they would point out the same problems as we have discussed in this thesis without influencing their answers. Possibly, we were hoping for some new problems to which we have not paid enough attention. For these interviews, the service providers were targeted.

The second way of gathering data was in the form of a survey with concrete questions to guide the respondents. Questions were divided into three parts: technical, security related, and data classification questions. The target group for the survey were companies that have recently implemented cloud computing services, or are currently in the process of implementation. That way they would have had recent experience with the problems they had to encounter in order to make the implementation work.

5.2 Interviews



	IBM	Citrix Systems
Cloud computing exposure	++	++
Complexity of security prerequisites	-	++
State of security solutions	++	-
Data classification necessity	+	++
State of data classification solutions	N/A	++

Figure 9: Representation of respondents' opinions to topics in conducted interviews. Source: Own compilation[7].

[7] The answers are based on a subjective analysis of the respondents' answers. Thus, if the respondent's opinion to a topic was highly positive, he was awarded "++"; if it was rather positive, a "+"; and if it was rather negative, a "-" was given.

Cloud computing exposure

From the respondents' answers[8] (see also Figure 9), it is apparent that companies which provide IT services are not ignoring cloud computing. In fact they are developing their own ranges of solutions to meet various customers' demands.

Complexity of security prerequisites

In his answer to the question "What are the (security) prerequisites for a company to begin using cloud computing", Mr. Martin Jelok, a Server and Storage Services Manager of IBM noted that his company currently concentrates mostly on private and so-called hybrid clouds. This suggests that IBM is aware of the risks a company faces in the cloud, but they are probably higher than the solutions. For this reason it currently recommends private clouds where operations take place "*in a 'defined' (secure) environment*" and reduce the need of complex security measures.

Mr. Volker Schmeisser, Country Manager Eastern Europe Central of Citrix Systems also talks about the prerequisites for using public clouds. In his opinion, specific security issues which should be raised before using cloud computing are:

- Privileged user access
- Regulatory compliance
- Data location
- Data segregation
- Recovery
- Long-term viability

We have talked about all these factors in chapters 2 and 3 in this thesis. In a comment to data segregation, Mr. Schmeisser claims: *"Data in the cloud is typically in a shared environment alongside data from other customers. Encryption is effective but isn't a cure-all."* He basically refers to the need of data classification to improve the decision making of handling data. In the Conclusion we made a similar remark, stating that no matter how much assurance we can get from the service providers, some data will never be placed in a (public) cloud.

State of security solutions

A major drawback for Mr. Schmeisser is that *"to date no official cloud-security industry standard has been ratified"* and that usually an SLA must be agreed to without the possibility to modify single elements. This introduces *"privacy, regulatory, and reputation risks"*. Mr. Jelok only noted that

[8] See Appendix 1 for complete interviews.

"cloud security is much about virtualization security and network security". In our opinion this is true only if other factors are eliminated by the use of private clouds.

Data classification necessity

Both respondents agree that a data classification system is necessary for cloud computing, although Mr. Jelok doubted if automatic data classification is a necessity.

State of data classification solutions

Mr. Schmeisser thinks there is a wide range of sophisticated data classification solutions which should be satisfactory. What he sees as a problem and a possible cause of project failure though, is an unclear implementation concept, i.e. a classification strategy. In this thesis, sections 3.3 and 3.4 are devoted to this topic as we also find this essential.

5.3 Survey

The target group for the survey[9] were companies that have recently implemented cloud computing services, or are currently in the process of implementation. The goal of the survey was to confirm the studied issues of implementing cloud computing and also to investigate how these companies are aware and prioritize various risks associated with the process.

The first part of the survey was a technical introduction. The results show that not everyone is absolutely sure about the differentiation between SaaS, Paas and IaaS. Mr. Snead from Qualcomm, a communication service provider, noted that there still might be some costs associated with the service which he might not be aware of. The negotiation of the SLA was very complex for the respondents, which confirms its importance and at the same time shows that these companies concentrate on it with a great deal. Nevertheless, the whole process of cloud service implementation was relatively short and took about 6 months. The customers usually intend to use it regularly for storage and archiving, but also for specific applications like CRM systems or customer facing applications.

The second part of the survey was about security. With the exception of the theatre, assuming which does not store a lot of confidential data, the results confirm the survey of Symantec (discussed in section 2.3.3) which said that companies face serious cyber attacks on a regular basis. At the same time the respondents claim that a service disruption of their cloud computing provider would have almost serious consequences. Surprisingly, although these companies had to severely rethink their

[9] See Appendix 2 for complete results and tables with input data.

data storage policies, they feel quite confident with the security at all levels of communication with the servers.

The final part of the survey was regarding data classification. Again, similarly to the interviews, approached companies think that a data classification system is a necessity for cloud computing, although not all of them are convinced if work towards classification automation is important. A possible explanation for this might be that companies have not really thought about this option, because they have not encountered such an autonomous system so far. Another explanation might be that because these companies are experimenting with cloud computing only partially, implementing automatic data classification into their IS might be too complex and unnecessary at the moment.

5.4 Conclusion

From the findings we can conclude that studied issues in this thesis are current and relevant. Clients of cloud computing services are careful with their security and are still not completely convinced with the security solutions that the service providers offer. The approached providers, on the other hand, seemed quite confident. Due to the clients' doubts, they tend to use cloud computing only partially for some projects, or they prefer internal clouds which ensure a more closed secure environment. Overall, the process of implementing a cloud computing project is very complex, requires the rethinking of data storage policies and various compliance standards need to be explored. Nevertheless, it is possible to implement a project successfully within six months.

6 : Conclusions

6.1 Conclusions

Cloud computing is without a doubt a hot topic within the business world. Since it is a relatively new technology, its implementation requires careful consideration. The intention of this thesis was to discover and examine these factors. We also examine some solutions to determine if they are sufficient for any company to begin using cloud computing. We came to a conclusion that automatic data classification is a necessary condition for a successful full implementation, therefore we elaborated on this topic. We observed some inefficiencies and outlined a solution that should solve them. By answering the sub questions first, we give more background to the answers of the main research question.

SQ1: What technical possibilities do cloud computing services offer?

Before determining the factors that need to be considered when implementing cloud computing, we need to understand the technical possibilities this technology can offer. Since cloud computing is a new concept, a certain level of confusion used to rule over the definitions, which has more or less settled nowadays and we can say that we generally differentiate between three major scenarios: Software-as-a-Service, Platform-as-a-Service and Infrastructure-as-a-Service. The differences are in the level of services which range from hosting web applications online, through a platform where non-web applications are hosted and run, to a whole IS infrastructure with storage and almost limitless number of virtual computers accessible online. Like with everything, all concepts have their advantages and disadvantages and are suitable for different tasks.

SQ2: What are the security concerns of companies' information systems?

Security is the first and foremost drawback slowing down the adoption of cloud computing. It is a distributed system, which have been used for as long as we have the internet. Therefore, security concerns in the communication paths between the clients and servers are well understood and solutions are available from the cloud computing predecessors.

The idea of a public cloud is to share hardware resources with other clients in order to save running costs. This is the cause of most of the problems, because storing data on a hard disc which is accessible by potential competitors, moreover not being the owner of this infrastructure can cause a lot of concerns. Another problem is that no standard has been ratified for cloud computing providers. This way it is difficult to satisfy various compliance standards which were not written with cloud computing in consideration. Since there will always be a certain level of risk of data leakage in the cloud, some kinds of data are simply not suitable to be sent there. Data owners need to be aware of the location of the files, who has access to them and if they should satisfy any compliance standards.

SQ3: What are the possible security measures to address the security concerns?

The confidentiality, integrity and availability of data can be maintained through cryptography, secure channels, access control, key management, authorization, and secure group management. The negotiation of the SLA can help solving compliance, liability and other legal issues. It can be also used to determine the countries in which the data will be physically stored.

A data storage policy determines what data can be sent to the cloud and how it should be protected. For this reason it is necessary to develop a data classification strategy. All respondents of the empirical study agreed with this fact. During the process of the strategy formation, it should be made clear what kinds of data flow within the organization and how these kinds should be treated and allowed access to the cloud. During the preparation of this descriptive study, we have encountered only partial technical solutions by various vendors. These might be sufficient for certain types of organizations or cloud computing projects, but they are incomplete for a full cloud computing integration.

SQ4: How necessary is automatic data classification when considering using cloud computing services?

Although some vendors and users of cloud computing are not convinced that further work towards data classification automation is necessary, this thesis tried to show that the opposite is true. It is not only recommended by security experts, but the opinion is also supported by facts like the ever increasing amounts of data being created, and that the way data is being stored is usually ineffective. Moreover, switching to an improved data classification strategy and applying it on the vast amounts of data already created would be a daunting task without the help of some automation.

We do not recommend a fully automatic system, but it should be as autonomous as possible. This is acceptable, because the types of data which flow within an organization keep reoccurring and computers should be able to learn how to classify them and build classification models from their knowledge. Technical solutions like Data Loss Prevention systems are not autonomous, because they only work with a predefined set of rules and exceptions. An aggressive classification parameter setting would increase the time administrators would have to spend reviewing each event. On the other hand, an autonomous system would first try to classify the data based on previous experience with such files possibly using a certain level of prediction, and only then if unsuccessful, it would request human assistance.

SQ5: How to classify data automatically?

To be able to classify data automatically, the computer needs to base its decision on some semantic information, which is referred to as metadata. First of all we have to differentiate between various types of data. Most of it is unstructured, but there is also a certain amount of structured data in the form of databases for example. Anyway, a single approach to all data is not possible. For example, textual and multimedia files clearly need a different approach to metadata extraction.

In this thesis we have presented a concept that describes the components such an automatic data classification system should contain. It is a combination of various metadata extraction and analysis methods. For example, extracting metadata from textual files is based on the study of Data Analysis and Recognition and multimedia files can be analyzed using an ISO standard MPEG-7. To ensure the autonomy of the classification system, incremental algorithms are required; therefore methods of machine learning and data mining are proposed. This way in theory, an autonomous and flexible system should be able to classify data automatically with little user interaction.

The main research question of this thesis stated the following:

RQ: What are the current roadblocks for companies considering cloud computing services?

There is no off-the-shelf solution for every company which decides to use cloud computing, but everything needs to be tailored to the customer's needs. The factors that need to be explored before a cloud computing implementation can be divided into several categories:

- *Compliance, data storage policy and legal.* Compliance and legal issues are considered to be one of the biggest roadblocks for companies considering cloud computing services. Most of them can be overcome by a comprehensive SLA with the service provider. In this agreement, the parties should clear out issues of for example intellectual property rights, privacy laws, liabilities and taxation. This way cloud computing users are able to protect themselves legally when they comply with various standards and regulations. The companies are also required to rethink their data storage policies. The rules and decision making should be based on the value of the data and should address issues of confidentiality, privacy, availability, ownership and distribution, integrity, data retention, and auditability.
- *Security;* is another factor blocking a fast adoption of cloud computing. From this perspective, data sent and stored in the cloud needs to have the assurance of

confidentiality, integrity and availability. Fortunately, the technical solutions that can give reasonable assurance of these factors are well developed. To be able to execute the data storage policy, the company's IS is in need of a data classification system.

- *Risk.* Relying solely on cloud computing can be dangerous in case of a service outage. This roadblock will be eliminated completely only when a cloud computing standard will be ratified, allowing for a simple recovery and migration of data between various service providers. The risk of service outage could then be solved by the redundancy of service providers.
- *Pricing.* Users of cloud computing services pay per use. This means they are charged precisely for things like the amount of processing power they consume, data they store and bandwidth they use. For this reason, some applications may not save costs by migrating to the cloud, especially when they create a lot of traffic. Either way, the pricing should be negotiated together with the SLA.
- *Latency.* Applications that require an extremely low latency between the decision and execution like for example high frequency trading systems are also unsuitable for the cloud.

The bottom line is that cloud computing is a great way to use hardware resources effectively and economically. From the findings we can conclude that it might be a great choice for deployment of some new projects. On the other hand, because of the fact that currently not all of the factors have been solved completely, we do not recommend using cloud computing as a full replacement for a physical IT infrastructure. Investing in cloud computing services should be carefully analyzed and must not be forced just to follow some trends. Hopefully, this thesis highlights all major factors that need to be considered before such a decision and can serve as a guide for further research.

6.2 Future research

During the process of writing this thesis, some questions emerged which need more research, but were out of the scope of this thesis. These are the most important future research questions:

FR1: How to standardize cloud computing datacenters?

The fact that to date, no international cloud computing standard has been ratified is the main cause of security concerns for its users. Almost each service provider stores data in a different way, therefore it is very difficult to migrate data from one provider to another. A standardization would amongst others solve the risk of a provider ending its services and guarantee a certain level of security.

FR2: Investigate compliance issues under cloud computing.

Almost all companies deploying a project in the cloud are struggling to satisfy various compliance and policy standards. Is it possible to satisfy the compliance standards in their current form, or is there a need for their novelizations? What needs to be done so that companies can avoid using internal clouds and enjoy all cloud computing benefits by sharing a public cloud? Further research from an audit and legal point of view is necessary to answer these questions.

FR3: Is it possible to build an automatic data classification system which would work for all types of data?

One of the reasons why we do not recommend using cloud computing as a full replacement for an IS infrastructure, is that we are not aware of a data classification system which would understand all types of data. The storage and movement of data within a company infrastructure is usually not effective enough to be shifted to the cloud. In this thesis we have described components such a data classification system should contain, but no attempts have been made to build and test it.

Bibliography

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of the Twentieth International Conference in Very Large Data Bases* (pp. 487-499). San Francisco: Morgan Kaufmann.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning* , 37-66.

Appleyard, J. (2003). Information Classification: A Corporate Implementation Guide. In H. F. Tipton, & M. Krause, *Information Security Management Handbook, Fifth Edition* (pp. 715-726). Auerbach.

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., et al. (2009). *Above the Clouds: A Berkeley View of Cloud Computing.* Berkeley: University of California.

Barsky, N., & Catanach, A. (2010, March 20). Letters. *Economist* , p. 18.

Bean, L. (2009). Cloud Computing: What Internal Auditors Need To Know. *Internal Auditing* , 34-38.

Belapurkar, A., Chakrabarti, A., Ponnapalli, H., Varadarajan, N., Padmanabhuni, S., & Sundarrajan, S. (2009). *Distributed Systems Security: Issues, Processes and Solutions.* John Wiley & Sons, Ltd.

Brodkin, J. (2008, July 22). *More outages hit Amazon's S3 storage service.* Retrieved July 24, 2010, from Computerworld: http://www.computerworld.com/s/article/9110463/More_outages_hit_Amazon_s_S3_storage_service

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems 25* , 599-616.

Clarke, S. (2002, May 24). *Creating a data storage policy.* Retrieved April 20, 2010, from ComputerWeekly.com: http://www.computerweekly.com/Articles/2002/05/24/187390/creating-a-data-storage-policy.htm

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* , 103-130.

Economist. (2010, May 15). Spotting video piracy: To catch a thief. *Economist* , p. 84.

Esposito, F., Ferilli, S., Basile, T. M., & Di Mauro, N. (2008). Machine Learning for Digital Document Processing: from Layout Analysis to Metadata Extraction. *Studies in Computational Intelligence* , 105-138.

Etges, R., CISA, CISSP, & McNeil, K. (2006). Understanding Data Classification Based on Business and Security Requirements. *ISACA Journal Vol. 5* , 1-8.

Execution MiH. (2010). *Metadata Management definition - What is metadata?* Retrieved May 11, 2010, from Execution MiH: http://www.executionmih.com/metadata/definition-concept.php

Google. (2009). *Google Labs*. Retrieved June 1, 2010, from Google Similar Images: http://similar-images.googlelabs.com/

Grimes, S. (2008). *Unstructured Data and the 80 Percent Rule*. Retrieved May 15, 2010, from Clarabridge Bridgepoints:
http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551

Hey, T., & Trefethen, A. (2003). *The Data Deluge: An e-Science Perspective.* UK e-Science Core Programme.

Hobson, D. (2009). Into the Cloud We Go... *Cloud Computing Journal Vol. 2, Issue 3* , 8-9.

Holte, R. C. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning* , 63-90.

Hunter, J. (1999, September). *D-Lib Magazine*. Retrieved May 31, 2010, from MPEG-7 Behind the Scenes: http://www.dlib.org/dlib/september99/hunter/09hunter.html

Chang, S.-F., Sikora, T., & Puri, A. (2001). Overview of the MPEG-7 Standard. *IEEE Transactions on Circuits and Systems for Video Technology* , 688-695.

ISO. (2002). *International Organization for Standardization*. Retrieved May 31, 2010, from ISO/IEC 15938-1:2002:
http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34228

John, G. H., & Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 338-345). San Mateo: Morgan Kaufmann.

Kshemkalyani, A. D., & Singhal, M. (2008). *Distributed Computing: Principles, Algorithms, and Systems.* Cambridge University Press.

Laird, P. (2009, May 19). *Cloud Computing Taxonomy at Interop Las Vegas, May 2009*. Retrieved May 17, 2010, from Laird OnDemand: http://peterlaird.blogspot.com/2009/05/cloud-computing-taxonomy-at-interop-las.html

Laird, P., & Oracle. (2008, May 29). *SaaS Soup: Navigating the "as a Service" Acronyms: CaaS, DaaS, DBaaS, PaaS, SaaS, XaaS*. Retrieved May 17, 2010, from Laird OnDemand: http://peterlaird.blogspot.com/2008/05/saas-soup-navigating-a-service-acronyms.html

Leffel, C. (2010, February 17). *Introduction to Data Loss Prevention*. Retrieved May 14, 2010, from TrueDLP: http://www.codegreennetworks.com/blog/?p=33

Maloof, M. A. (2006). *Machine Learning and Data Mining for Computer Security.* London: Springer.

Maloof, M. A. (2006). Some Basic Concepts of Machine Learning and Data Mining. In M. A. Maloof, *Machine Learning and Data Mining for Computer Security* (pp. 23-43). London: Springer.

Marinai, S. (2008). Introduction to Document Analysis and Recognition. *Studies in Computational Intelligence* , 1-20.

Marinai, S., & Fujisawa, H. (2008). *Machine Learning in Document Analysis and Recognition.* Berlin: Springer.

Masinter, L., Archer, P., Connolly, D., Hammer-Lahav, E., & Denenberg, R. (2001). *Metadata on the Web: A survey*. Retrieved May 11, 2010, from w3.org: http://www.w3.org/2001/tag/2009/02/metadata-survey.html

Minsky, M. L., & Papert, S. (1988). *Perceptrons : An Introduction to Computational Geometry.* Cambridge: MIT.

Nilsson, M., Palmér, M., & Naeve, A. (2002). SemanticWeb Meta-data for e-Learning – Some Architectural Guidelines. *11th World Wide Web Conference (WWW2002)* (pp. 1-22). Hawaii: Royal Institute of Technology.

Parrilli, D. M. (2010). Legal Issues in Grid and Cloud Computing. In K. Stanoevska-Slabeva, T. Wozniak, & S. Ristol, *Grid and Cloud Computing: A Business Perspective on Technology and Applications* (pp. 97-118). Berlin: Springer.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Mateo: Morgan Kaufmann.

Raval, V., & Fichadia, A. (2007). *Risks, Controls and Security: Concepts and Applications.* John Wiley & Sons.

Rosenblum, M. (2004). The Reincarnation of Virtual Machines. *Queue Volume 2 , Issue 5* , 34-40.

Schuller, S. (2008, December 1). *Demystifying The Cloud: Where Do SaaS, PaaS and Other Acronyms Fit In?* Retrieved May 17, 2010, from SaaS Blogs: http://www.saasblogs.com/2008/12/01/demystifying-the-cloud-where-do-saas-paas-and-other-acronyms-fit-in/

Stanoevska-Slabeva, K., Wozniak, T., & Ristol, S. (2010). *Grid and Cloud Computing: A Business Perspective on Technology and Applications.* Berlin: Springer.

Symantec. (2010). *2010 State of Enterprise Security Report.* Symantec.

Takebayashi, T., Tsuda, H., Hasebe, T., & Masuoka, R. (2010). Data Loss Prevention Technologies. *Fujitsu Science Technology Journal, Vol. 46, No. 1* , 47-55.

The Economist. (2010, March 13). Amazon auctions computing power. *The Economist* , p. 64.

Toigo, J. W. (2005). *Data Classification in Distributed Computing Environments: Getting off to a Good Start.* Dunedin: Toigo Partners International.

Vaquero, L., Rodero-Merino, L., Caceres, J., & Lindner, M. (2009). A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review 39, no. 1* , 50-55.

Vellante, D. (2007, May 2). *Data classification value transcends storage efficiencies.* Retrieved May 9, 2010, from Wikibon: http://wikibon.org/wiki/v/Data_Classification

Vellante, D., & Floyer, D. (2008, June 3). *Data classification: Brains or brawn?* Retrieved May 9, 2010, from Wikibon: http://wikibon.org/wiki/v/Data_Classification

Velte, A. T., Velte, T. J., & Elsenpeter, R. (2010). *Cloud Computing: A Practical Approach.* McGraw-Hill.

Vouk, M. A. (2008). Cloud Computing – Issues, Research and Implementations. *Journal of Computing and Information Technology Vol. 16, No. 4* , 235-246.

Wendt, J. M. (2010, April 7). *DLP's Data Classification and Ownership Features do not Mask the Value of Inactive Data, They Uncover It*. Retrieved June 25, 2010, from DCIG: http://symantec.dciginc.com/2010/04/dlp-does-not-mask-value-inactive-data.html

Williams, A. (2010). *Finance Midmarket Client Unstructured Data Analysis Report.* Nottingham: Centiq.

YouTube. (2010). *Content Management*. Retrieved May 31, 2010, from YouTube: http://www.youtube.com/t/contentid

Zirnoon, T. (2009, January 27). *What is the difference between DLP, ILP, CMF and EPS ?* Retrieved May 13, 2010, from McAfee Security Insights Blog: http://siblog.mcafee.com/data-protection/what-is-the-difference-between-dlp-ilp-cmf-and-eps/

Appendix 1

Interviews are based on email communication with respondents.

Respondent 1 (June 11th 2010):

Martin Jelok
Global Technology Services,
Server & Storage Services Manager
IBM Slovensko, spol. s r.o.

Respondent 2 (June 16th 2010):

Volker Schmeisser
Country Manager Eastern Europe Central
Citrix Systems

Q: Does your company have any experience with implementing cloud computing services / Did it encounter such services? Is your company considering using them?

Mr. Jelok: Globally yes, it is our hot topic. In Slovakia - not yet.

Mr. Schmeisser: Citrix Systems is offering cloud computing solutions. These solutions are powered by an open and extensible suite of technologies. These technologies enable a range of solutions, from on boarding application workloads to the cloud, to optimizing application performance, to effectively managing towards enterprise SLAs—all while delivering on the promise of cloud economics, elasticity, and scale. Extensibility and on-demand customization are built into the design of each Citrix Cloud Solutions technology component to address the unique needs of every enterprise. A high degree of interoperability among Citrix cloud technology components and third party technologies ensure that there is no lock-in when customers adopt a Citrix Cloud Solution. Enabling customer choice is a fundamental goal of the underlying technology powering Citrix Cloud Solutions.

Q: In your opinion, what are the (security) prerequisites for a company to begin using cloud computing? What should it be especially careful about?

Mr. Jelok: IBM at the moment invests in private clouds - where clouds operate in a "defined" (secure) environment. Public clouds and hybrid cloud solutions are on the roadmap. Some public clouds are already available.

Mr. Schmeisser: Generally it can be stated that there are specific security issues customers should be raised before using cloud computing.

Privileged user access.
Sensitive data processed outside the enterprise brings with it an inherent level of risk, because outsourced services bypass the "physical, logical and personnel controls" IT shops exert over in-house programs. Get as much information as you can about the people who manage your data.

Regulatory compliance.
Customers are ultimately responsible for the security and integrity of their own data, even when it is held by a service provider.

Data location.
When you use the cloud, you probably won't know exactly where your data is hosted. In fact, you might not even know what country it will be stored in.

Data segregation.
Data in the cloud is typically in a shared environment alongside data from other customers. Encryption[10] is effective but isn't a cure-all.

Recovery.
Even if you don't know where your data is, a cloud provider should tell you what will happen to your data and service in case of a disaster[11].

Long-term viability.
Ideally, your cloud computing provider will never go broke or get acquired[12] and swallowed up by a larger company.

Q: Do you think the current state of security solutions is sufficient for cloud implementation? Where do you see drawbacks?

Mr. Jelok: IBM cares about security and I believe IBM cloud solutions currently on the market have sufficient security in-built. Security solutions are evolving as fast as other technology is evolving. They always come hand-in hand. Cloud security is much about virtualization security, network security. There are security solutions on market to meet cloud computing security needs.

[10] http://www.networkworld.com/news/2008/022108-disk-encryption-cracked.html
[11] http://www.networkworld.com/topics/backup-recovery.html
[12] http://www.networkworld.com/slideshows/2008/mergers-and-acquisitions.html

Mr. Schmeisser: A large drawback is the fact that to date no official cloud-security industry-standard has been ratified. Each could provider takes a different approach to security and most of them are not forthcoming about their security architectures and policies. Due to missing standards, security appears to be just another SLA. In many cases the cloud provider's standard SLA must be agreed to (or disagreed to) as a package, without offering the possibility to modify single elements in order to fit the cloud consumer's security needs. However, viewing cloud security as an SLA introduces privacy, regulatory, and reputation risks. If prospect cloud consumers are considering the cloud for security, they should ask there provider the following questions:

1. What happens when the SLA fails?
2. How are policy and regulatory compliance enforced?
3. How can geographic data distribution be controlled?
4. Who reports mandatory data breach notifications?
5. Who pays for data breaches?
6. What is the opportunity cost of reputational risk?
7. How are physical audits and forensic analysis performed?

Q: In your opinion, is a data classification system a necessity for cloud computing? Can you think of other technical systems?

Mr. Jelok: You better contact my colleague Viktor Homolya, on cc: I think data classification is necessary, but not sure if automated data classification is a necessity. It is also a question whether (and what) data is on (public) cloud.

Mr. Schmeisser: When data resides outside the company premises it is ever more important to segregate public from sensitive information. Data classification is therefore a most essential measure for secure cloud computing. Understanding where Personally Identifiable Information, Personal Healthcare Information, Payment Card Industry information, and Intellectual Property reside helps enforce governance and regulations.

Q: Are you satisfied with the current state of data classification solutions? What is missing?

Mr. Jelok: Please ask Viktor Homolya[13].

Mr. Schmeisser: There is a large range of sophisticated data classification solutions on the market today, and with a clear implementation concept in place there is no reason why a data classification project should fail. The lack of a good concept (what data should be classified with which level, and

[13] Mr. Homolya was contacted but has not found the time to reply.

for how long, etc.) is more often the cause of unsuccessful projects than missing product features. Often the nature and amount of sensitive data is not correctly assessed before launching the project, or department heads and teams are not being involved sufficiently in the up-front decision making.

Appendix 2

Information was collected through a temporary online survey on www.thesistools.com.

Automatic data classification for cloud computing security

Dear Sir/Madam,

Please take a few minutes to fill out this survey for my thesis on the topic "Automatic data classification for cloud computing security".

Target group:

Companies which have recently implemented cloud computing services, or are currently in the process of implementation.

Goal of the thesis:

Discover factors which need to be explored in order to securely implement Cloud Computing Services (CC) in a company

Goal of this survey:

- Confirm studied issues of implementing CC with CC customers
- Investigate how companies prioritize the risks (awareness)

Please fill out the following fields:

Reference	Name	Name of company	Position within company	Email contact
A	Chris Chatfield	Qualcomm	Information Architect	cchatfie@qualcomm.com
B	John Snead	Qualcomm	Staff Systems Analyst	jsnead@qualcomm.com
C	Fumi Matsumoto	Allurent	Co-founder and CTO	fmatsumoto@allurent.com
D	Igor Kocina	Theatre Srećko Kosovel	PR	igor.kocina@gmail.com

Technical part

Instructions

The questions are written with the assumption you have already finished negotiating with the service provider(s). In case you have not finished yet, please try to answer the questions as if they would be in the present form.

	A	B	C	D
How well do you think you understand the differentiation between SaaS, PaaS and IaaS? Are the possibilities of cloud computing clear to you? (1 - Confused, 5 - Clear)	3	5	5	2
Were all the pricing and running costs explained to you sufficiently? (1 - Not at all, 5 - Sufficiently)	0	4	4	5
How complex was the negotiation of the SLA? (1 - Fast, 5 - Complex)	0	5	4	4

How do you intend to use cloud computing at your company?

	A	B	C	D
Internally (Internal cloud)	Yes	Yes	No	No
Seasonally, sporadically	No	No	No	No
Host some SW online in order to reduce number of SW licenses	No	No	Yes	No
Storage, archiving	No	Yes	Yes	Yes
Data mining	No	No	No	No
Other (please specify)	Internal and External CRM and Knowledge Base	Customer Facing Applications living outside corp firewall	A SaaS offering of our interactive merchandising capabilities for online retailers	

How long did the whole process of cloud computing services implementation take? (From the moment of decision to first live runs)

A	6 months
B	6 months
C	4 months
D	6 months

Security part

	A	B	C	D
How often does your company experience serious cyber attacks? (1 - None, 5 - Very often)	4	3	3	1
How much would a cloud computing service disruption affect you? (1 - Not at all, 5 - Serious consequences)	4	4	4	4
Did you have to rethink your data storage policies? (1 - Not at all, 5 - Completely)	4	4	4	4

How concerned are you with the security at the level of the (1 – Concerned, 5 – Confident):

	A	B	C	D
Service provider	4	4	4	4
Data transfer (Communication with server)	4	4	4	4
Client side (Your company, end users)	4	3	4	5

Please list some compliance standards you had to explore before you could implement CC services.

B	Legal Agreements with customers, data storage and access security policies, SAS70
C	PCI DSS, data storage policies
D	None

Data classification part

Data classification is a part of the Information Lifecycle Management (ILM) process. Its goal is to collect certain information about data to improve the decision making of its use. Data classification should determine how sensitive certain data is, where it is located, who has access to it etc. It could be a manual process, but could be also automated to some extent. For example, Data Loss Prevention (DLP) systems can be considered as some kind of a data classification system.

	B	C	D
How important do you consider a data classification system due to cloud computing service implementation? (1 – Unnecessary, 5 – Essential)	4	4	4
How important is it to work towards data classification automation? (1 – Unnecessary, 5 – Essential)	3	4	5
Did your company have a data classification system before you were considering cloud computing?	Yes	Yes	Yes
Does your company have a data classification system now?	Yes	Yes	Yes

If you answered "No" to the previous question, you can skip the remaining questions on this page and submit the survey.

	B	C	D
How difficult was it to construct a data classification strategy? (1 - Fast, 5 - Labor intensive, Considerable time)	4	5	4
How complex do you think your strategy is? (1 - Basic rules, 5 - Very complex, many exceptions)	3	4	2

How much attention does your data classification strategy give to the following types of data? (1 – Not important, 5 – Plays a crucial role)

	B	C	D
Structured data (databases)	4	4	4
Emails	4	3	2
"Digitally-born" documents (textual files, spreadsheets, presentations, etc.)	4	3	4
Scanned documents (e.g. in form of PDF)	2	3	1
Multimedia files (video conference archive, telephone conversation, video presentation, etc.)	4	2	1

How well does your implemented system solve these classification strategies? (1 – Not supported, 5 – Fully satisfied)

	B	C	D
Structured data (databases)	4	4	4
Emails	4	4	2
"Digitally-born" documents (textual files, spreadsheets, presentations, etc.)	4	4	4
Scanned documents (e.g. in form of PDF)	4	4	1
Multimedia files (video conference archive, telephone conversation, video presentation, etc.)	3	3	1