TALLINN UNIVERSITY OF TECHNOLOGY

School of Information Technologies

Külli Kool 183175IABM

Data security analysis for cloud service in Organization X

Master's thesis

Supervisor: Jaak Tepandi

Professor Emeritus

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Külli Kool 183175IABM

Andmeturbeanalüüs pilvepõhise keskkonna jaoks Organisatsiooni X näitel

Magistritöö

Juhendaja: Jaak Tepandi

Emeriitprofessor

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Külli Kool

10.05.2021

Abstract

An organization preferring to remain anonymous (referred to in the thesis as Organization X), is planning to switch to cloud services, but it does not have internal data security regulations for setting the appropriate data security levels and measures.

The goal of the current master thesis was to carry out data security analysis based on metadata fields of information assets in Organization X. The broader purpose of the thesis was to create a data security evaluation list based on metadata fields, which would simplify the decision-making process when publishing data to cloud services.

Data security analysis was performed in accordance with the three – level baseline security system ISKE (ISKE) [1]. The first step was an information assets inventory - all digital forms with metadata fields were described and divided into categories based on their content. Then, security subclasses were assessed by content categories and security classes were formed. Security levels were assessed based on security classes and limitations for cloud services were determined.

The problem was that all metadata fields had to be evaluated one by one for achieving the information asset security level. The Apriori algorithm was used as an association rules mining technique for exploring the connections and their strength between metadata fields with the intention to use the discovery for simplifying the security level assignment process.

Combinations of metadata fields, which always occur together on digital forms, were detected by association rules mining, and previously assessed security levels were added. The outcome of this (presented in Appendix 2) was the data security evaluation list based on combinations of metadata fields, which can be used for assessing and assigning security levels and measures for cloud services. It could be used on already existing and newly created metadata fields and it can be valuable input as pre-analysis for developing data security evaluation tool based on metadata fields.

Metadata fields differ between organizations, but each organization can create their own data security evaluation list by using the same techniques for content analysis: data security level assessment by ISKE and association rules mining by the Apriori algorithm.

This thesis is written in English and is 68 pages long, including 5 chapters, 12 figures and 8 tables.

Annotatsioon

Andmeturbeanalüüs pilvepõhise keskkonna jaoks Organisatsiooni X näitel

Organisatsioon, kes avaldas soovi jääda anonüümseks (Organisatsioon X), kavandab pilveteenustele üleminekut, kuid puudu on sisemised eeskirjad andmete turbeastmete ja -meetmete määramiseks.

Käesoleva magistritöö eesmärk oli viia läbi asutuse infovarade metaandmete väljadel põhinev andmeturbeanalüüs Organisatsioonis X. Lõputöö laiemaks eesmärgiks oli luua metaandmeväljadel põhinev andmeturbe hindamise loend, mis lihtsustab andmete pilveteenustesse avaldamise otsustusprotsessi.

Andmeturbeanalüüs tehti vastavalt "INFOSÜSTEEMIDE KOLMEASTMELISE ETALONTURBE SÜSTEEMI ISKE rakendusjuhendile" (ISKE) [1]. Esmalt viidi läbi infovarade inventuur ning kirjeldati kõik digitaalsed vormid koos metaandmeväljadega. Metaandmeväljad jaotati kategooriatesse sisu alusel ning hinnati turvaosaklassid, mille alusel moodustati turvaklassid. Nende põhjal omakorda määrati turbeaste ja piirangud pilveteenuste jaoks.

Probleemiks oli, et infovara turvalisuse taseme saavutamiseks tuleks kõiki metaandmevälju ükshaaval hinnata. Turbeastme määramise protsessi lihtsustamiseks otsustati kasutada Apriori algoritmi, mis on assotsiatsioonireeglite kaevandamise tehnika. Peamine eesmärk oli uurida metaandmeväljade vahelisi seoseid ja nende tugevust.

Metaandmeväljade kombinatsioonid, mis esinevad digitaalsetel vormidel alati koos, tuvastati assotsiatsioonireeglite kaevandamise teel, ja lisati varem hinnatud turbeastmed. Tulemuseks (Lisa 2) oli metaandmeväljade kombinatsioonidel põhinev andmeturbe hindamise loend, mida saab kasutada pilveteenuste turbeastmete ja -meetmete hindamiseks ja määramiseks juba olemasolevate ja vastloodud metaandmeväljade puhul. Lisaks võib see olla väärtuslik eelanalüüs metaandmeväljadel põhineva andmeturbe hindamise tööriista väljatöötamisel.

Metaandmeväljad on organisatsioonides erinevad, kuid iga organisatsioon saab koostada oma andmeturbe hindamise loendi, kasutades andmete analüüsiks samu võtteid: andmete turvalisuse taseme hindamine ISKE alusel ja assotsiatsiooni reeglite kaevandamine Apriori algoritmiga.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 68 leheküljel, 5 peatükki, 12 joonist, 8 tabelit.

List of abbreviations and terms

CIA	The three most important properties of data - confidentiality, integrity and availability (CIA).
Cloud computing	The delivery of different IT services through the Internet. This includes data storage.
Cloud service	Infrastructure, platforms, or software that are hosted by third-party providers and made available to users through the internet.
Data	Set of values of qualitative or quantitative variables about one or more objects.
Data sensitivity	Concerns information that should be protected from unauthorized access due to its sensitive nature.
Digital form	The electronic version of a paper form, which is accessible in any location.
Information asset	Information or data that is valuable to the organization.
Information life cycle	The stages every item of information goes through from its creation to its final archiving or destruction.
Information security	The practice of preventing unauthorized management of information.
Information system	A technical system processing, storing or transmitting data, along with the means, resources, and processes needed for its normal operation [2, p. 17].
ISKE	The three-level IT baseline security system of information systems [2, p. 17].
ISMS	Information security management system.
Metadata field	Input elements that are populated with data to describe information assets.
Security class	Shows the criticality of data, expressed on a four-level scale and with three components, i.e. as a combination of three security subclasses [2, p. 18].
Security level	Information security indicator assigned based on a security class. ISKE includes three security levels L – low, M – medium, and H – high [2, p. 18].

Security measure	Organisational acts and means, technical processes and implementation of technical means for obtaining and retaining the safety of data and data in information systems [2, p. 19].
Security subclass	Level required to obtain the purpose of information security based on the criticality of data, expressed on a four-level scale [2, p. 19].

Table of contents

List of figures

List of tables

1 Introduction

In the scope of this master thesis, an organization data security analysis based on information assets metadata fields was carried out and data security levels for cloud services were assessed according to the three – level baseline security system ISKE (ISKE) [1]. The main purpose of the thesis was to create a data security evaluation list based on metadata fields used in Organization X, which would simplify the decision-making process, when publishing data to cloud services.

Chapter 2 gives an overview of previous related research and the current situation in the organization X.

The data security analysis, assessment of security levels and limitations for cloud services based on ISKE is described in chapter 3.

The data security analysis continues in chapter 4 with association rules mining for discovering patterns in information assets, assigning security levels to results and forming the final result.

1.1 Problem description

Organization X was established in the second half of 20th century with a mission to construct and operate metro systems. Today, it has approximately 5000 employees who are responsible for more than one million trips carried out with metro, commuter and express trains.

One of its head offices with about 200 employees is located in Europe and it is a parent company to 4 subsidiaries. The duties of its Business Strategy Department include IT management. The Business IT Department manages Microsoft SharePoint software for company´s Intranet. The SharePoint software is mainly used for creating and managing websites for different suborganizations, but also for storing, organizing and accessing documents, lists, calendars and other information securely from all devices via a web browser [3].

Cloud computing has grown in popularity and cloud storage has become one of the hotspots of information storage [4]. Organization X is also planning to start using cloud-based technology, including SharePoint Online, a cloud-based service hosted by Microsoft which allows to create sites for sharing information and documents with internal and external customers [3].

Privacy and security are the biggest concerns in cloud storage [4]. Data segregation, protection and leak prevention are considered the most important security challenges in cloud based services [5]. This has highlighted a problem regarding the use of SharePoint Online in the future: the company doesn't have enough knowledge and rules about content security in the cloud service.

Cloud security can be divided into cloud data security and storage security. Data security ensures the privacy of data and storage security ensures the correctness of data uploaded to the servers [6]. Since Organization X is planning to migrate to cloud services and does not have any internal regulations for data security, the thesis will focus on data security.

1.2 Thesis goal

The goal of the thesis was to address the problem that the company does not have sufficient knowledge and established rules about external and internal content security in cloud service, as well as find an answer to the question of how to categorize the digital content by sensitivity level and decide which can be migrated online and which not.

Today, most of the information in Organization X is created and managed digitally in different information systems and preserved in company´s own servers. The organization does not have previous experience with storing data and documents in the cloud, neither do they have any rules or internal regulations for information security in regard to this.

The organization's main aim is to avoid risks related to sensitive data leakage in the cloud-based service. That is the reason the company would like to carry out information assets analysis starting from Intranet. Plan is to divide information assets into categories and assign security levels based on international frameworks. On the basis of security levels, the company can decide, which information and how it is reasonable to manage through the information life cycle in SharePoint Online cloud service.

1.3 Expected results

As a result of information assets analysis, all information assets created in the Intranet would be described along with their metadata fields. Metadata fields will be divided into categories based on their content; security level will be assessed and assigned to each metadata field content category. The highest metadata security level will be assigned to information asset.

The expected outcome of this master thesis is a data security evaluation list which could be used for evaluating data security based on metadata fields. With the help of this list, an organization could assign a security level to every piece of existing or created data in Intranet. Once the security level was defined, it would be possible to decide whether the data should be stored on a cloud service (and if so, on which conditions), or in an on-premises system with higher security requirements.

Metadata fields differ between organizations, but each organization can create their own data security evaluation list by using the same techniques for content analysis: data security level assessment by ISKE and association rules mining by Apriori algorithm.

Designing a technical solution for the protection of the data according to its security level is a topic for further exploration.

1.4 Methodology

The methodology used in this thesis relied on data security analysis based on "Implementation manual for the Three-level baseline security system ISKE" [2] and association rules mining by the Apriori algorithm as the basic and most influential technique in the association rule mining area. [7, p. 1].

All information assets (digital forms) were exported from the current Intranet along with metadata fields and divided into categories based on the metadata field content. Then, security subclasses of each content category were estimated and a security class was assigned to the metadata field content category. Following that, the security level of the metadata field content category was assessed according to security class based on instruction for using cloud services [8]. The highest security class of digital form metadata fields was assigned to the digital form itself.

Next, association rules mining was used to discover some hidden patterns, which could simplify the assignment of security levels to digital forms and their metadata. For that purpose, the Apriori algorithm in RStudio was used [9]. Then it was investigated whether and how it was possible to assign security levels based on detected connections between metadata fields.

The expected outcome of the analysis was the list of metadata field combinations along with security levels, which can be used for evaluating and assigning data security levels.

2 Description of current situation

Chapter gives the overview of the related research on cloud computing and data security issues and describes the current information system in Organization X.

2.1 Previous research

Cloud computing has grown in popularity and along with that, cloud storage has become one of the hot topics of information storage. Data security is one of the most important problems related to cloud storage that needs to be addressed urgently [4].

The risks of data storage are mainly related to two aspects:

1. Hardware devices and applications. Users' data is stored in different storage spaces in complex multi – user environments and there are no backups. It is highly possible that user data can be intercepted, damaged or lost if the service provider doesn´t encrypt data or apply recovery and backup functions.

2. The service provider. Cloud storage devices are managed by third party and users do not know how the data is stored. Service providers get control over data after it´s uploaded and they have the responsibility to provide a trusted service and ensure security of data - the question is how users can verify this [4].

Organizations are gaining more experience in cloud services and they start to shift business functions to cloud platforms. It´s a growing trend for companies around the world. Some organizations go along with this, while for others the idea of storing sensitive data outside the premise is unimaginable [10].

A company should study their processes and evaluate advantages and risks before making the shift to the cloud. The biggest advantage of cloud storage is reduced cost, but there are also several risks, including security and privacy concerns. It is not certain whether the cloud computing model adequately protects sensitive information [10].

The cloud computing domain consists of several subdomains; each of these could have different security and privacy requirements. There are several issues to focus on - authentication and identity management, access management etc, but the core issue is privacy and data protection. Many companies are hesitant to store their data in the cloud because there´s a risk of unauthorized access. Service providers must embed security solutions to all their services and offer high transparency security policies [11].

Security challenges need to be addressed, before implementing cloud computing in the organization. Data operations and transmissions are at high risk if the security measures are not applied properly. Measures must be taken to identify and handle challenges. Data privacy and security is the most critical factor to consider. Data loss or leakage can have a huge impact on the business functions and consumer´s trust for the whole organization [5].

Encryption is suggested as one the best solutions for securement. Before uploading data to the server, it should be encrypted. Also, key management techniques can be used to protect against unauthorized access [5].

Data leakage or data corruption can lead to people's mistrust and the collapse of a company. Data security issues can be divided into four categories:

1. CIA-related security issues. The three most important properties of data are confidentiality, integrity and availability (CIA). These must be preserved during the data life cycle, which has six stages: create, store, use, share, archive and destroy.

2. AAC-related security issues. Authentication and Access Control (AAC) must be applied not only to people but also to machines. It is the process of verification and authentication to connect and get access to cloud resources.

3. Broken authentication and session control. These threats can occur if user credentials are insufficiently protected or there is a lack of restrictions for what is allowed for authenticated users.

4. Other risks. Reasons related to data location, multi – tenancy, backups in cloud etc [12].

Encryption is seen as the most appealing solution to security problems, but it requires expensive infrastructure for implementation [13].

The suggested security protection levels are:

1. Protected. Applied to data for public or free distribution. Data not critical to user needs - for example, marketing materials. Usually password protected.

2. Sensitive. Classified as medium sensitivity data, for non-public views, usually business data. Loss or detriment of this data does not have severe impact. Protected by multi-factor authentication.

3. Top secret. Restricted or confidential data, the loss of which would have catastrophic consequences. For example, personal data that enables identification. Full encryption is suggested as a protection method [13].

The security level may change and vary per type of data or per demand of data owner. This is the reason for identification, because the suitable encryption method could be suggested based on that. The application of encryption technique does not always ensure the most efficient security and there´s a need for a dynamic and more effective data security system [14].

Based on the three most important properties of data: confidentiality, integrity and availability (the CIA triad), some methods are suggested for data security:

1. Data security flow chart. Following the yes and no questions will help the user to identify the data type and its security level.

2. Data Security based on weights of CIA parameters. Information sensitivity categories: low, medium high, are correlated with CIA parameters in the cross table. User can decide, which class the data belongs to and apply relevant protection methods.

3. Sensitivity rating. The assessment of CIA triad parameters could be done automatically based on the impact caused by the combination of these parameters. Combination of these three parameters is called sensitivity rating and it can be formulized: SR = C + I + A. The ratio of each parameter varies.

4. Different methods for SR calculation based on inversely proportional values [14].

Data security includes the identification of data elements by their value in the business. Data can be classified by three types of characteristics:

1. Access control. Defines access restrictions to data and includes frequency of access and updates, visibility and accessibility, retention period.

2. Content and its properties: accuracy, reliability, degree of completeness, consistency, auditability.

3. Storage and its policies based on criteria applied to different data types: storage and communication encryption, access control, recovery and backup, quality standards [15].

This set of parameters can be used for data security in the cloud and it provides security levels based on content type. All data stored in the cloud is first classified by content and then by restriction levels. The resulting assessment can be used as security provisions for storage and communication encryption, integrity and access control mechanisms [15].

Ensuring the safety of cloud storage is a complex issue. While service providers claim data stored securely, companies remain hesitant. A framework based on Service level agreements (SLA) should ensure data security in cloud storage. SLA is an agreement between the service provider and customer and it defines the service type, quality and payment terms [16].

There are three different technologies to keep data safe in the cloud:

1. Storage protection. Data is divided into small pieces and stored in different locations. If one data centre or disk crashes, then data can be still restored.

2. Transfer protection. As storage devices are located at a considerable distance from the customer, the data must be transferred through the network using proper security protocols.

3. Authorization. User authority and access control is important and all operations must be recorded and traceable [16].

Cloud security can be divided into cloud data security and storage security. Data security ensures the privacy of data and storage security ensures the correctness of data which is uploaded to the servers [6].

Data security can be classified into three categories:

1. Preservation of privacy. Ensuring the privacy of personal and business-critical information is a crucial problem and requirements to the confidentiality and authorization must be established.

2. Storage security. Integrity must be preserved as it is one of the critical properties of data.

3. Data security. It is a process of protecting data from unauthorized use and establishing a policy for protecting sensitive information [6].

Lack of knowledge in data security can lead to a critical error in business functions, which may result in financial loss [17]. A security assessment system used for traditional information systems is not suitable for cloud, because cloud computing uses virtualization technology and cloud security has become the key limitation to the further implementation of cloud computing [18].

In conclusion, most solutions to privacy and security issues focus on technical solutions related to encryption, which usually requires expensive infrastructure for implementation. Frameworks based on security for storing data in the cloud have been less researched. The goal of the thesis is to help to bridge the gap and create a data security evaluation list based on an international framework to simplify an organization's decision-making process when publishing content to a cloud service.

2.2 Current information system

Microsoft SharePoint is used for creating websites. The information can be stored, organized, shared and accessed from different devices via the web browser [3].

Organization X has been using Microsoft SharePoint since 2012. They started with SharePoint 2007, an upgrade to SharePoint Foundation was done in 2015. SharePoint

Foundation is only available for SharePoint Server 2013 and it can be used for creating collaborative web pages, various documents, lists, calendars and information types [3].

The organization´s intranet has been running on the SharePoint on – premise platform and it consists of two site collections: intranet and workplaces.

Intranet pages include news, blogs, general information about company and tools for employees, which help them organize everyday work.

In the workplace site collection, each subcompany has its own page for aggregating different work-related content about the subcompany and its functions.

Organization X plans to move to SharePoint Online, which is a powerful cloud-based application used for empowering teamwork, quick information retrieval and seamless collaboration across the company [19].

For achieving this goal, the organization must migrate most of the existing content to cloud-based SharePoint Online. Before, it is necessary to carry out analysis for determining which content and on which conditions, is safe to publish to cloud.

The current solution contains 20 digital forms, which are used in SharePoint applications. Below is the list of forms with a short description:

- Sales Deviation Report. A form for reporting the difference between the income and ticket sale numbers.
- Cleaning items orders. A form for ordering new cleaning items to the station.
- Inventory material in service counter. Information regarding inventory of materials in the service counter.
- Emergency evacuation command test. Information about upcoming tests.
- Private exchange. Application for a work shift with a colleague.
- Station control night. Information about performed station checks.
- Fault report. A report about a traffic disruption with a duration over ten minutes.

- Time assignments. Information about occasions when vacancy staff has to allocate overtime.
- Traffic incident report. Must always be written in the event of threats/violence and security situations.
- Improvement suggestion. A form for proposing work related improvements.
- Maintenance book. A form for vehicle maintenance materials.
- Event calendar. Calendar for different events.
- Transport calendar. Calendar about vehicle transportation.
- Vacation application. A form for applying a vacation.
- Task management. A form for assigning and managing tasks.
- Result card. For associating company's overall goals with employee´s individual goals and activities. These consist of physical cards that indicate the most important goals from the balance sheet for the year.
- User target. A form for employee goals, related to result card application.
- Company target. A form for company goals, related to result card.
- Contact list. For employee contact information.
- Gold Grains. A form for appreciating colleagues.

Each form has create, edit and display view and metadata on these can be different. Metadata is structured information which enables the management of documentation through time and across domains. It can be used to identify, authenticate and contextualize documents, people, functions and systems, which create, manage, preserve and use them [13, p 2]. Its main purpose is to support business and records management processes [13, p 3].

Data processed in this system is related to organization´s support and core functions and is mainly:

- Personal data, which enables to identify a person directly or indirectly by an identifier such as name, identification number, location data, an online identifier or by physical, physiological, genetic, mental, economic, cultural or social identity factors [21]. In the current application, these are: first and last name, employment number, department, position, boss, phone number.

- Financial and assets management data. A source accounting document must as minimum contain information about the economic transaction: its date, content, numerical values (amount, price and total sum) [15, ch 2]. Current application includes information about sales, orders, assets.

- Trains' and stations' management information, which consists of several vehicle and station management procedures.

- Employee management related data, which includes information for stations' and train management, but also information related to human resources processes.

All digital forms have an internal security level according to the Organization X document management manual. It means that this information may not be disclosed to third parties without the permission of the responsible publisher. Information is accessible to all employees, including rental workforce, but use of the company's devices is recommended, device must be registered and meet the IT department's security requirements [23].

3 Data security analysis based on ISKE

Software quality is a degree to which a system can be used to satisfy its stakeholders stated and implied requirements [17, p 17]. The international standard ISO/IEC 25010:2011 “Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models” defines a product quality model, which provides a framework for quality requirements and evaluation [17, p 19]. It´s a defined set of product quality properties categorized into eight characteristics and security is one of them [17, p 10].

Security is a degree to which a system protects data so that persons, other products or systems have access according to their types and levels of authorization [17, p 13]. Related sub-characteristics are confidentiality, integrity, non-repudiation, accountability, authenticity [17, p 10]. It means that only authorized persons have access and modification permissions to data, it is possible to prove that actions or events have taken place and they are traceable and can be proved to be the one claimed [17, p 14].

There are also other information security related standards. International standard ISO/IEC 27001:2013 „Information technology. Security techniques. Information security management systems. Requirements“ belongs to the ISO/IEC 27000 “Information Security Management Systems” standard family. This family consists of inter-related standards and guidelines, which help to manage the security of organization information assets [25].

ISO/IEC 27001:2013 provides requirements for an information security management system (ISMS), which is used for managing information securely. ISMS must be part of an organization´s processes and information security must be considered in the design of processes and systems [18, p v].

ISO/IEC 27002:2017 “Information technology. Security techniques. Code of practice for information security controls“ is designed for organizations, who are implementing information security management controls and guidelines It establishes the guidelines and

general principles for initiating, implementing, maintaining, and improving information security management in an organization [26].

The Control Objectives for Information and related Technology (COBIT) is published by the Standards Board of Information Systems Audit and Control Association (ISACA). COBIT 5 is a business framework for the governance and management of company IT including the information security part [27].

The Information Technology Infrastructure Library (ITIL) is a collection of best practices in IT service management (ITSM) that focuses on the service processes of IT and considers the central role of the user. It also includes IT security management [28].

Additional cloud-specific guidance based on ISO/IEC 27002 is provided by ISO/IEC 27017:2021 „Information technology. Security techniques. Code of practice for information security controls based on ISO/IEC 27002 for cloud services“. It addresses cloud-specific information security threats and provides measures [29].

ISKE [1] is the three – level baseline security system based on the IT Baseline Protection Manual (IT-Grundschutz) published by the German BSI (Federal Office for Information Security) [30]. The purpose of its implementation is to ensure a sufficient level of security for the data processed in information systems [1]. It also has a complete implementation guide [31] and instruction for a safe use of cloud services [8].

The author will focus on ISKE [31] in the first step of data sensitivity analysis.

3.1 Assessment of data security

ISKE is meant for achieving and maintaining the security of information assets and is applicable both in the private and public sectors [31, p. 3].

It provides three security levels:

- Low (L)
- Medium (M)
- High (H) [25, p 4]

Organization X plans to upgrade their SharePoint Foundation 2013 on-premise platform to SharePoint Online and was advised to consider ISKE requirements before a new information system will be implemented or the currently existing system will be improved. [29, p 4]. The focus was on the migration to cloud service and for achieving this goal, the author assigned security levels to all currently existing digital forms in SharePoint Foundation 2013 and made suggestions regarding the secure use of cloud services [8]. The author and the organization´s SharePoint manager are responsible for this assignment.

3.1.1 Information assets inventory

According to ISKE, the first step is to carry out an information assets inventory and [31, p. 5]. It is recommended to have minimal information about each component in the separate table, but the information must be detailed enough for the implementation of ISKE [31, p. 7].

All currently existing information assets in SharePoint were described in an Excel table with following columns (an example is presented in Table 1):

- Information_Type. Here it´s a *digital form* in all cases.
- Information_Type_Name. The column contains the names of 20 digital forms
- Information_Type_Form. Contains new, edit and display form values, because metadata can be different on those forms.
- Field_Name. Contains the field names of 939 metadata fields.

Information_Type	Information_Type_Name	Information_Type_For	Field_Name
Digital form	Sales Deviation Report	New form	Has there been a plus or minus at the checkout
Digital form	Sales Deviation Report	New form	My first and last name
Digital form	Sales Deviation Report	New form	My employment number

Table 1 Information assets described to Excel file

It´s advised to group similar information assets to simplify the management of ISKE and implementation of security measures [31, p. 8]. The “Content” column was added to the Excel table and an example is given in Table 2.

Information_Type	Information_Type_Name	Information_Type_For	Field_Name	Content
Digital form	Sales Deviation Report	New form	Has there been a plus or minus at the checkout	financial data
Digital form	Sales Deviation Report	New form	My first and last name	personal data
Digital form	Sales Deviation Report	New form	My employment number	personal data
Digital form	Sales Deviation Report	New form	My immediate boss	personal data
Digital form	Sales Deviation Report	New form	I belong to area	personal data

Table 2 Information assets with data content category column

All digital forms metadata fields were divided into seven categories by their content: assets management, employee management, financial data, general management, personal data, station management, train management. This is given in the "Content" column in Table 2.

3.1.2 Assessment of security classes

The next step was to assign security classes according to ISKE [31, p. 5]. ISKE uses a security model which relies on the CIA triad and refers to the main elements of security measures in information systems: availability, integrity and confidentiality [32, p. 22].

This concept was created in 1970s and after 50 years it´s still uniquely important in security management [32, p. 29].

Availability is the actual availability of organization´s data to authenticated persons and systems. All access channels must work properly to ensure the protection and availability when needed. It is the primary requirement of each information system [31, p. 9].

Integrity protects data from unauthorized modification or deletion and that damage made by authorized person can be reversed [31, p. 9].

Confidentiality means that data is accessible only to authorized persons or systems and inaccessible to all others. An organization must be capable of defining and enforcing access levels for information [31, p. 9].

The required information security level depends on the organization's functions, regulations and internal instructions etc. Data security means that the CIA triad is achieved [31, p. 9].

ISKE uses a four-level scale for assigning security level and relies on CIA. By applying a four-level scale to CIA, it is possible to determine security subclasses, which consist of a security goal symbol and value [31, p. 9]:

Availability:

K0 - less than 90% per year; the maximum acceptable single interruption may exceed 24 hours.

K1 - greater than or equal to 90% and less than 99% per year and maximum acceptable single interruption is up to 24 hours.

K2 - greater than or equal to 99% and less than 99.9% per year. Maximum acceptable single interruption is up to 4 hours.

K3 - greater than and equal to 99.9% per year. Maximum acceptable single interruption is up to 1 hour [31, p. 9].

Integrity:

T0 – information source, identifiable disposal or modification is not important. Information accuracy, integrity and timeliness checks are not required.

T1 - information source, disposal and modification must be identifiable. Information accuracy, integrity and timeliness checks in specific cases and as required.

T2 - information source, disposal and modification must be identifiable. Periodical checks of information accuracy, integrity and timeliness are required.

T3 - information source, disposal and modification must have probative value. Real - time verification of information accuracy, integrity and timeliness are required [31, p. 10].

Confidentiality:

S0 - public information, access to information is not restricted.

S1 - information for internal use, access to the information is allowed in the case of a legitimate interest.

S2 - confidential information, the use of information is allowed only for certain groups of users, or in the case of a legitimate interest.

S3 - top secret information, the use of information is allowed only for certain specific users or in case of a legitimate interest [31, p. 10].

For assessing the security class of the digital forms fields on the basis of the four-level scale described above, columns for security subclasses named "Availability (K)", "Integrity" (I) and "Confidentiality" (S) were added to the table; an example is provided in Table 3.

The following requirements must be considered when determining security subclasses:

- Requirements from legal acts and agreements.
- Requirements from organization business functions.
- Assessment of consequences:
    - R0 – no remarkable damage caused by a security incident.
    - R1 – unimportant damage, remarkable restrictions to organization core functions or financial loss.
    - R2 – important damage and restriction for executing core processes
    - R3 – critical damage and core functions not completed, critical financial loss [31, p. 11].

These requirements were met by adding the following columns to Excel (example given in Table 3):

- "Avail_Regulations" for assessing the impact of legal acts and agreements to availability. Assessment was based on Estonian government regulation "The system of security measures for information systems" [33].
- "Avail_Core_FN" for assessing the impact of internal regulations to core functions to availability. Assessment was based on organization X document management manual [23].

- "Avail_Conseq" for assessing the impact of security incident consequences to availability. Assessment was based on ISKE [31, p. 11].

- "Integ_Regulations" for assessing the impact of legal acts and agreements to integrity. Assessment based on the Estonian government regulation "The system of security measures for information systems" [33].

- "Integ_Core_FN" for assessing the impact of internal regulations to core functions to integrity. Assessment based on the Organization X document management manual [23].

- "Integ_Conseq" for assessing the impact of security incident consequences to integrity. Assessment was based on ISKE [31, p. 11].

- "Conf_Regulations" for assessing the impact of legal acts and agreements to confidentiality. Assessment based on Estonian government regulation "The system of security measures for information systems" [33].

- "Conf_Core_FN", assessing the impact of internal regulations to core functions to confidentiality. Assessment based on organization X document management manual [23].

- "Conf_Conseq" for assessing the impact of security incident consequences to confidentiality. Assessment was based on ISKE [31, p. 11].

Content	Avail_Regulations	Avail_Core_FN	Avail_Conseq	Availability (K)	Integ_Regulations	Integ_Core_FN	Integ_Conseq	Integrity (T)	Conf_Regulations	Conf_Core_FN	Conf_Conseq	Confidentiality (S)	Security_Class
assets management	1	2	0	**2**	1	1	0	**1**	1	1	0	**1**	**K2T1S1**
employee management	2	2	1	**2**	2	2	2	**2**	1	1	2	**2**	**K2T2S2**
financial data	2	2	0	**2**	2	2	2	**2**	1	1	2	**2**	**K2T2S2**
general management	1	1	0	**1**	1	1	0	**1**	1	1	0	**1**	**K1T1S1**
personal data	2	2	1	**2**	2	1	2	**2**	1	1	1	**1**	**K2T2S1**
station management	3	3	3	**3**	3	3	2	**3**	1	1	3	**3**	**K3T3S3**
train management	3	3	3	**3**	3	3	3	**3**	1	1	3	**3**	**K3T3S3**

Table 3 Data content categories with security subclasses and classes

Each requirement was assessed and the result added to proper column. The highest requirement value determined the security subclass, Availability (K), Integrity (T) or Confidentiality (S), value. It is presented in Table 3.

First, the "assets management" content category was evaluated. The impact of legal acts, internal regulations and consequences to "Availability" was assessed and it was

discovered that internal regulations have the strongest impact according to organization X document management manual [23]. The highest value – “2” - was assigned to “Availability” and it means that the maximum acceptable service interruption can be up to 4 hours [31, pp. 9–10].

Next, integrity was evaluated for the “assets management” category. The value is “1” according to laws and internal regulations and it means that the information source, disposal and modification must be identifiable and checkable in specific cases and as required [31, pp. 9–10].

Confidentiality was determined to be “1” according to laws and internal regulations and it means that this information is meant for internal use only.

Next, the data security class was assigned to the “assets management” metadata fields content category. Data security class is a combination of three security component classes, the values of Availability (K), Integrity (T) and Confidentiality (S) and the maximum number of combinations is 64. The data security class marking is always formed based on the markings of subclasses in their order K-T-S. [31, p. 10]. Security classes are given in “Security_Class” column in Table 3.

The value determined for availability was “2”, for integrity “1” and for confidentiality “1”, resulting in security class **K2T1S1** for “assets management“ category as seen in “Security_Class” column in Table 3.

Assessment of data content categories “employee management” and “financial data” both resulted in security class **K2T2S2**. Availability and integrity for this data was estimated to be “2”, because of the requirements from external and internal regulations. The consequences of confidentiality were set to “2” because of important loss it may cause [31, p. 11].

General management data security class was assessed to **K1T1S1**, because of low requirements of laws and internal regulations. Also the consequences do not cause important damage or financial loss [31, p. 11].

Personal data has availability valued to “2”, which means maximum allowed service downtime up to 4 hours according to external and internal regulations [31, pp. 9–10].

Integrity is also “2” because of important damage it may cause to organization and stricter requirements from external regulations. Confidentiality result is “1” because this data is restricted for internal use only. The security class **K2T2S1** was formed.

Train and station management data is related to organization core functions and is the most critical, which is also concluded from its security class **K3T3S3.** It has the highest requirements for availability, where maximum downtime up to 1 hour is required. The integrity value requires real - time verification of information accuracy and timeliness [31, pp. 9–10]. It also has the highest confidentiality requirements because of consequences, which may end up with critical damage [31, p. 11].

3.1.3 Assessment of security levels

Security level is assigned to data according to security class marked in “Security_Class” column in Table 3.

In Figure 1 below are given three ISKE baseline security levels matched to 64 security class combinations:

- Low security level (L),
- Medium security level (M),
- High security level (H) [31, p. 14]

		K0	K1	K2	K3
T0	S0	L	L	M	H
	S1	L	L	M	H
	S2	M	M	M	H
	S3	H	H	H	H
T1	S0	L	L	M	H
	S1	L	L	M	H
	S2	M	M	M	H
	S3	H	H	H	H
T2	S0	M	M	M	H
	S1	M	M	M	H
	S2	M	M	M	H
	S3	H	H	H	H
T3	S0	H	H	H	H
	S1	H	H	H	H
	S2	H	H	H	H
	S3	H	H	H	H

Figure 1 ISKE security classes and levels

The author had previously divided organization content into seven categories in "Content" column in Table 2 Information assets with data content category column On the basis of these categories, security classes were assigned to data in Table 3. Next, security levels were defined by Figure 1 and presented in column "Data_Security_Level" in Table 4 below.

Content	Security_Class	Data_Security_Level
assets management	K2T1S1	M
employee management	K2T2S2	M
financial data	K2T2S2	M
general management	K1T1S1	L
personal data	K2T2S1	M
station management	K3T3S3	H
train management	K3T3S3	H

Table 4 Data content category security classes and levels

At first, the security level for data content category "assets management" was assessed. It has security class **K2T1S1** in "Security_Class" column. First symbol "T" was used and it refers to security component "Integrity" and it´s value is "1". The security subclass value "S" refers to "Confidentiality" and it´s "1". The last symbol, "K" stands for "Availability" with a value of "2". Consequently, the security level according to Figure 1 ISKE security is "M" - medium.

Data content categories "employee management" and "financial data" both have security class **K2T2S2** and referring to security level "M", medium.

Data content category "general management" has security class **K1T1S1.** It has the lowest security subclass values and therefore also the low (L) security level.

Personal data has security class **K2T2S1** matching the medium (M) security level.

Data content categories "Station management" and "Train management" have the highest security subclass values in **K3T3S3** and therefore also the highest (H) security level.

3.2 Assigning limitations for cloud service

Instruction for using cloud services securely has been created based on ISKE [8]. Its purpose is to describe different cloud services and to explain why and which security measures must be implemented [26 p2]. It focuses on Estonian public sector organizations and the current master thesis attempted to implement it in organization X, because it´s similar to Estonian public sector organizations in its structure and internal regulations.

Requirement to use of cloud services is to assign ISKE security classes to data which will be processed in cloud service. Main restrictions come from security subclasses [8, p. 10].

ISKE gives the main rules and restrictions for maintaining data in cloud service by security subclasses. These are presented in Figure 2 below. L, M or H stands for ISKE security level [8, p. 10].

				KÄIDELDAVUS			
				K0	K1	K2	K3
TERVIKLUS	T0	KONFIDENTSIAALSUS	S0	L	L	M	H
			S1	L	L	M	H
			S2	M	M	M	H
			S3	H	H	H	H
	T1	KONFIDENTSIAALSUS	S0	L	L	M	H
			S1	L	L	M	H
			S2	M	M	M	H
			S3	H	H	H	H
	T2	KONFIDENTSIAALSUS	S0	M	M	M	H
			S1	M	M	M	H
			S2	M	M	M	H
			S3	H	H	H	H
	T3	KONFIDENTSIAALSUS	S0	H	H	H	H
			S1	H	H	H	H
			S2	H	H	H	H
			S3	H	H	H	H

Figure 2 ISKE limitations assignment for cloud services.

The meaning of the cell colours:

- Green – the use of the public cloud is allowed within the European Union
- Yellow - the use of the public cloud is allowed under certain circumstances.
- Orange - the use of the public cloud is allowed under certain circumstances, assessment of risks and backup of data to the system located in origin country.
- Red – the use of public cloud services is not allowed [8, p. 10].

Security level cell was coloured in Table 5 below according to Figure 2:

Content	Security_Class	Data_Security_Level
assets management	K2T1S1	M
employee management	K2T2S2	M
financial data	K2T2S2	M
general management	K1T1S1	L
personal data	K2T2S1	M
station management	K3T3S3	H
train management	K3T3S3	H

Table 5 Security levels for cloud services

Most of the security level cells are yellow in colour, which means that the use of the public cloud is allowed under certain conditions [8, p. 10]. These are determined by security subclasses [8, p. 11]

Confidentiality

S0 - data can be processed in a public cloud and stored in EU.

S1 and **S2** - use of public clouds is allowed, but data encryption is required at rest as well as in transit.

S3 - use of public cloud service is not allowed [8, p. 11].

Availability

K0 - data may be processed in a public cloud and stored in EU.

K1, K2 - use of public clouds is allowed, but the risks must be evaluated separately, and regular data backups must be done to another, independent system. Cloud service monitoring and incident management is mandatory.

K3 - in addition to K1 / K2 measures, backups must be done to another system located in origin country [8, p. 12].

Integrity

T1 - Normal integrity measures, such as access control and logs, must be implemented.

T2 - In addition to T1 measures, integrity logs must be sent to another system.

T3 - in addition to T1 and T2 measures, integrity logs must be sent to another system in origin country and integrity must be ensured with strong cryptography [8, p. 12].

Assets management related data has security class **K2T1S1** and a yellow security level. The security component class for Availability (K) is "2", which means that the risks of availability must be assessed separately and regular data backups must be made to other independent systems. Monitoring of cloud service and management of security incidents must be organized [8, p. 12].

Integrity component class value for assets management is "1", which means regular integrity assurance methods must be implemented, for example access permissions and activity logs [8, p. 12].

Confidentiality is "1", which means data must be encrypted at rest and in transit [8, p. 11].

Employee management related and financial data share the security class **K2T2S2** and the yellow medium security level. Requirements to availability are the same as for assets management data: the risks assessment, regular data backups and system monitoring [8, p. 12].

T2 for employee management and financial data requires the transfer of system logs to another system [8, p. 12]. S2 for confidentiality insists data encryption at rest and in transit [8, p. 11].

General management data has low yellow security level and security class **K1T1S1.** Availability conditions again require the separate risks assessment, regular data backups and system monitoring [8, p. 12]. The integrity subclass value calls for the implementation of regular integrity assurance methods, for example access permissions and activity logs [8, p. 12] and confidentiality requires data encryption at rest and in transit [8, p. 11].

Personal data has security class **K2T2S1** and yellow medium security level. The availability value requires that the risks of availability must be assessed separately and regular data backups as well as the monitoring of cloud service and management of security incidents [8, p. 12]. Integrity component class value "2" requires the transfer of system logs into another system [8, p. 12]. Confidentiality value requires data encryption at rest and in transit [8, p. 11].

Train and station management data has red high security level with a security class **K3T3S3.** The use of cloud services is not allowed for this data [8, p. 10].

It is possible to assign security level and its restrictions to a digital form based on the data content category security level.

In Table 6 digital form name is in column "Information_Type_Name" and security level is in column "Data_Security_Level" and it has proper restrictions colour [8, p. 10].

It is possible to see, which fields the digital form has and the field's security class in Excel table. It can be concluded that digital form security level is equal to its metadata highest security level. And in addition, restrictions applied to digital form are the highest restrictions applied to data the form includes.

Table 6 includes the list of digital forms with their security classes and levels for cloud service:

Information_Type_Name	Information_Type_Security_Class	Information_Type_Security_Level
Cleaning items orders	K2T2S2	M
Company Target	K2T2S2	M
Contatct List	K2T2S2	M
Emergency evacuation command test	K3T3S3	H
Event Calendar	K2T2S2	M
Fault report	K3T3S3	H
Gold Grains	K2T2S2	M
Improvement suggestion	K2T2S1	M
Inventory material in service counter	K2T2S2	M
Private exchange	K2T2S2	M
Result Card	K2T2S2	M
Sales Deviation report	K2T2S2	M
Station control night	K3T3S3	H
Task Management	K2T2S2	M
Time assignment	K2T2S2	M
Traffic incident report	K3T3S3	H
Transport Calendar	K3T3S3	H
User Target	K2T2S2	M
Vacation Application	K2T2S2	M

Table 6 Digital forms with security class and level for cloud services

It can be concluded that the following digital forms with a red high (H) security level must not be published to cloud service:

- Emergency evacuation command test
- Fault report
- Station control night
- Traffic incident report
- Transport Calendar

All other forms have medium yellow security level and security class **K2T2S2,** except the Improvement suggestion, which has security class **K2T2S1**.

According to ISKE, the highest security class and security measures must be applied to the whole system [31]. If digital forms with red high security level are excluded, not migrated, or will not be used in MS SharePoint Online, then the security level for whole system is yellow medium with security class **K2T2S2**.

It defines that the use of the public cloud is allowed under certain conditions for digital forms [8, p. 10]. Security component class for Availability (K) is “2” and it requires that the risks of availability must be assessed separately, and regular data backups must be

performed to other system. Monitoring of cloud services and management of security incidents is necessary [8, p. 12]. Integrity component class value is “2” , requiring the transfer of system logs into another system [8, p. 12]. Confidentiality component class value requires data encryption at rest and in transit [8, p. 11].

4 Data security analysis based on association rules

In the previous chapter, data security levels and measures for cloud services were assigned. A security level was assigned to each metadata content category separately. Herein lies the problem - all metadata fields had to be evaluated one by one to establish the information asset security level.

Each digital form consists of several metadata fields, which can repeat on different forms. Investigating the relationships between fields would help to seek patterns to simplify the security level assignment. Because of that association rules mining was used as one of the data mining techniques for determining the connection between metadata fields. Association rules mining is the most common data analysis technique, which has become a significant data mining solution [34, p. 1].

4.1 Data mining and association rules

Data science is an evolving discipline, which is defined as the process of extracting meaningful knowledge from data. It is often equated to data analytics, which creates models for problem solving. This process includes the following activities: data acquisition, pre-processing, modelling, testing and reporting the results [35, p. 4].

Data is acquired from Excel tables or streamed from databases, but models are built with the help of data mining [35, p. 4].

Data mining as a separate term was first used in the academic community in 1995. It has its origin in statistics, mathematics, machine learning, artificial intelligence, and business field. Data mining is defined as the process of finding patterns in data by using different algorithms and the outcome of the process is the generalized model of data with the goal to reuse discovered patterns in new situations [35, p. 5]. From the business perspective, it is important to carry out deeper analysis on a large amount of company data for discovering hidden rules to create models for supporting business activities. In short, data mining is the extraction of important knowledge from a data set [36, p. 1].

All data mining techniques are based on induction-based learning - forming general concept definitions by observing examples of concepts [35, p. 5]. A concept is a set of grouped objects which share specific characteristics and it´s the output of data mining session [29, p. 6].

The data mining process consists of the following repetitive steps:

- Mining objects determination. It is possible to perform data mining on all types of data sets.

- Data preparation, which includes data integration, selection, pre-processing and conversion. Integration covers solving different issues, for example semantic problems, missing data, cleaning data. After that, relevant data can be selected, pre-processed for overcoming data mining tool limitations and converted into a processable format.

- Data mining comprises selecting the proper data mining algorithm and performing the data mining.

- Results presentation and application - presenting the outcome of the analysis in an understandable form to users, evaluation of results, finding results and knowledge consolidation [36, p. 2].

One of most common data analysis techniques is association rule mining, which has become a significant data mining solution [34, p. 1]. It was first introduced by R.Agrawalet in 1993 [36, p. 2].

Association rule is a probabilistic rule which detects certain connections among a set of data items in the “if-then” statement form [34, p. 1]. It helps to discover relationships among transactions within a dataset and determine interesting connections between many items. Association rule data mining algorithm performance has impact on data mining efficiency, integrity and effectiveness [36, p. 1].

A formal description of association rule mining: assume that $I = \{i_1, i_2, \ldots, i_m\}$ is a set of elements, which consist *of m* different items. The database *D*, in which each transaction *T* is a set of items in *I*, that is $T \subseteq I$. If the item set is that $X \subseteq I$ and $X \subseteq T$, then the transaction *T* contains the item set *X* [36, p. 2].

Association rule is a conditional implication modelized as $X \Rightarrow Y$, in which $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \emptyset$ [34, p. 1]. X is called *antecedent* and *Y consequent* [37, p. 1].

A frequency and precision measure is associated to each rule for generating only the relevant relationships between the item sets [34, p. 1]. So, an association rule has two flag parameters:

- Support S - the database *D* should have at least *S%* transactions include all the items in *X* and *Y*;
- Confidence *C* - at least *C%* transactions including *X* contains *Y* [36, p. 2].

The criteria or evaluating the association rules is minimum support threshold (MST) and minimum confidence threshold (MCT). X ⇒ Yi is valuable if and only if support(X⇒Y) ≥ MST and confidence (X⇒Y) ≥ MCT [38, p. 2]. Minimum support and confidence thresholds are most commonly set by the domain experts and the value usually varies from 0% to 100%, not 0 to 1 [36, p. 2].

Apriori algorithm is the basic and most influential in association rule mining area. It was proposed by R.Agrawal in 1993 as an algorithm for mining single dimensional, single layer and Boolean association rules. The main idea was to use a recursive method of layer by layer search [7, p. 1].

The main idea of the algorithm is described as follows:

- First, scan the transaction database *D*, calculate the support of all 1- candidate itemset C_1 and compare with the minimum support. After that, generate the frequent l-item set L_1 which is not less than minimum support.
- Then L_1 join itself ($L_1 \infty L_1$) and prune it to generate the 2- candidates C_2. Select the item greater than minimum support to generate frequent 2-item sets L_2.
- After that, use L_2 to generate L_3; the process is repetitive and iterative, until the result meets the final qualification $L_3 = \phi$, the algorithm ends [33, p. 2].

There has been research into related data mining techniques, which focuses on big data mining security implications [39] and data mining techniques role importance in network

security [40]. Several works focus on privacy-preserving data mining techniques [41], securing the confidentiality of sensitive patterns for preserving privacy by hidden association rules [42] and privacy preserving in association rule mining by different algorithms [43].

The author focused on the most common algorithm in association rules mining, Apriori [44, p. 213]. The goal was to discover some unknown patterns in digital forms metadata fields which could help to assign data security levels by field connections and create a list for data security levels pre-evaluation.

4.2 Association rules mining by Apriori algorithm

The first step in data mining is the determination of mining objects. It is possible to perform data mining on all types of information storage, including relational and transactional databases, data warehouses etc [36, p. 2]. SharePoint uses SQL Server, which is a relational database management system [45]. Author exported all currently existing information assets from SharePoint and transformed them to an Excel table, which includes columns with the digital form name (Information_Type_Name) and the names of metadata fields belonging to digital form (Field_Name).

The next step is data preparation [36, p. 2]. Author selected relevant data, in current case all metadata fields of all digital forms. It was done based on the Excel table containing digital form and their metadata field names. Each digital form has new, display and edit views and metadata fields could repeat on these forms. The purpose of research was not to investigate the connections between digital forms and their views. Because of that, these relationships were considered not relevant. Consequently, the author focused on finding the associations of different metadata fields and removed all duplicated metadata fields related to one digital form. The outcome was the list of all digital forms and their unique metadata fields.

After that data was converted into a suitable format for processing. Digital form names were not relevant and were removed from the dataset. Each table row contains one digital form and includes all the metadata related to that exact form. Since 20 digital forms were investigated in this thesis, there were 20 rows with metadata fields (an example is given

in Table 7). “,” was used as a metadata field name separator and words in metadata field name were separated with “_”.

	A	B	C	D	E	F	G	H	I	J
1	My_first_and_last_name,My_employment_number,Station_entrance_and_service_desk,Station_area,C									
2	Title,Doc_number,Comments,Administrator,Approver,Published,Audience,Vehicle_type,Document_ty									
3	My_first_and_last_name,My_employment_number,I_belong_to_area,My_colleagues_first_and_last_na									
4	Customer_service_KPIs,My_goals,Individual_goal,Employee_name,My_immediate_boss,My_employme									
5	Has_there_been_a_plus_or_minus_at_the_checkout ,My_first_and_last_name ,My_employment_numb									
6	My_first_and_last_name,My_employment_number,Station,Station_area,Date_when_station_check_is_									
7	Assignment_number,Assignment_name,Start_date,Original_End_Date,End_date,Carriage_type,System,									
8	My_first_and_last_name,My_employment_number,I_belong_to_area,Time_information_refers_to_dat									
9	My_first_and_last_name,My_employment_number,My_position,I_belong_to,My_phone_number_priva									
10	Depot_from_and_to_and_wagon_type,Start_time,Select_a_date_in_the_calendar,End_time,From_and_									
11	KPI,Goal,Year,Department,Active									
12	Individual_goal,Employee_Name,My_immediate_boss,MTRS_Overall_Objectives,Goal,Year									
13	Name,Employment_number,Vacation_group,Vacation_with,Saved_vacation_days,Take_out_saved_wee									
14	Name,Position,Phone,Comment,E_mail,Manager,Company,Location_Department									
15	Barrier_line,Date_when_the_test_is_to_be_performed,Month,Service_center_has_called_station_host									
16	Heading,Place,Start_time,End_time,Description,All_day,Recurrent,Category,Event,Traffic_solution,Traf									
17	My_employment_number,Were_you_on_call_station_manager_during_the_disturbance,If_Yes_please_									
18	Heading,Motivation_for_the_Golden_Grain,Name_of_the_person_persons_you_want_to_give_a_Golde									
19	Heading,I_work_in,Name,The_improvement_pertains_to_companies_departments,The_improvement_									
20	My_first_and_last_name,My_employment_number,Date_and_time_when_inventory_is_performed,Sta									

Table 7 Example of pre-processed metadata fields

The data mining process includes selecting the proper data mining algorithm and performing the data mining [36, p. 2]. The author of the current thesis chose to use RStudio for the purpose. RStudio provides a free and open source integrated development environment for R, a programming language for statistical computing and graphics [9].

It is possible to use different RStudio packages for data processing. Package is a fundamental unit of reproducible R code. It includes reusable R functions and instructions with sample data [46]. There´s an arules package, which provides the infrastructure for representing, manipulating, and analysing transaction data and patterns by using frequent item sets and association rules. It also provides the association mining algorithm Apriori. Two arules core packages were used in the thesis:

- arules, a base package with data structures and mining algorithm Apriori.
- arulesViz, which helps to visualize the association rules [47].

Additionally, the RColorBrewer package was used, which provides different color palettes for graphics [48].

All these packages were installed, libraries and the data set were loaded. Transaction object was checked on in Figure 3:

```
> tr
transactions in sparse format with
 20 transactions (rows) and
 311 items (columns)
```

Figure 3 Transaction object check

There are 20 transactions and 311 items. That means 311 metadata field names and 20 collections of those.

A summary of the dataset is given in Figure 4.

```
> summary(tr)
transactions as itemMatrix in sparse format with
 20 rows (elements/itemsets/transactions) and
 311 columns (items) and a density of 0.05819936

most frequent items:
  My_employment_number My_first_and_last_name              Comments               Heading        Individual_goal               (Other)
                     9                       7                     3                     3                     3                   337

element (itemset/transaction) length distribution:
sizes
 5  6  8 10 11 12 15 21 27 34 35 40 43
 1  3  2  3  1  1  1  2  1  2  1  1  1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00    8.00   11.50   18.10   28.75   43.00

includes extended item information - examples:
                                                                                labels
1 A_green_brochure_with_the_text_To_you_who_have_been_in_a_serious_incident_on_the_front
2                                                 A4_paper_how_much_is_in_the_package
3                                                                              Active
```

Figure 4 Summary of dataset

Density gives the percentage of non-zero cells in sparse matrix [49]. It is possible to calculate, how many data fields are on digital forms in total by using density: 20x311x0.05819936 = 362.

The summary shows the most frequent datasets and element length distributions. It represents the number of transactions for 1-itemset, etc. The first number is the number of items and the second is showing transactions. The longest transaction has 43 items. Transactions numbers are low and item sets with 6 and 10 items have the highest transaction number – 3.

An item frequency plot was generated to visualize the distribution of the objects in Figure 5:

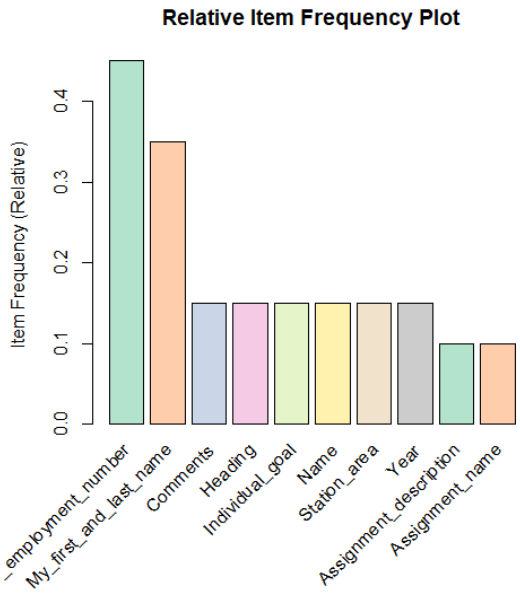

Figure 5 Relative item frequency plot

10 highest frequency items were plotted and it can be seen that the metadata fields *My employment number* and *My first and last name* appeared most frequently. The *My employment number* field appeared 0.5 times and *My first and last name* 0.35 times more frequently than other fields. It means that these fields are represented on most of the digital forms. Then frequency drops more than halves and other fields are not so common on digital forms.

The next step was to find interesting relationships between these 311 items and apriori algorithm is loaded for that. At first, default parameter values were used for mining: minimum support 0.1, confidence 0.8 and maximum of 10 items (maxlen).

Support shows the frequency of itemset occurrence and how significant it is [50]. The value 0.1 tells that 10% of all digital forms must include antecedent and consequent. Value 0.8 for confidence shows the probability of 80% seeing the rule´s consequent under the condition that the transaction also contains the antecedent. By default, maximum 10 items are presented in relationship.

```
> #rule 1
> apriori(tr,parameter = list(support = 0.1,confidence = 0.8)) -> rule1
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.8    0.1    1 none FALSE            TRUE       5     0.1      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 2

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[311 item(s), 20 transaction(s)] done [0.00s].
sorting and recoding items ... [33 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.00s].
writing ... [109415 rule(s)] done [0.03s].
creating S4 object  ... done [0.06s].
```

Figure 6 First results of Apriori algorithm

109 415 rules were given in Figure 6 and from Figure 7 it´s seen that the 7 and 8 item rule length has the most rules and length of 2 items have the lowest number of rules.

```
> summary(rule1)
set of 109415 rules

rule length distribution (lhs + rhs):sizes
    2     3     4     5     6     7     8     9    10
  190  1113  4008 10010 18018 24024 24024 18018 10010

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   6.000   7.000   7.316   9.000  10.000

summary of quality measures:
    support         confidence    coverage         lift            count
 Min.   :0.10   Min.   :1    Min.   :0.10   Min.   : 2.222   Min.   :2
 1st Qu.:0.10   1st Qu.:1    1st Qu.:0.10   1st Qu.:10.000   1st Qu.:2
 Median :0.10   Median :1    Median :0.10   Median :10.000   Median :2
 Mean   :0.10   Mean   :1    Mean   :0.10   Mean   : 9.761   Mean   :2
 3rd Qu.:0.10   3rd Qu.:1    3rd Qu.:0.10   3rd Qu.:10.000   3rd Qu.:2
 Max.   :0.35   Max.   :1    Max.   :0.35   Max.   :10.000   Max.   :7

mining info:
 data ntransactions support confidence
   tr            20     0.1        0.8
```

Figure 7 Summary of Apriori algorithm

Since there were 109 415 rules, then 10 of them, presented in Figure 8, were investigated further to have a first closer look.

```
> inspect(head(rule1,10))
     lhs                                  rhs                          support confidence coverage lift      count
[1]  {Goal}                            => {Year}                       0.1     1          0.1       6.666667 2
[2]  {My_immediate_boss}               => {Year}                       0.1     1          0.1       6.666667 2
[3]  {My_immediate_boss}               => {Individual_goal}            0.1     1          0.1       6.666667 2
[4]  {Station_entrance_and_service_desk} => {My_first_and_last_name}   0.1     1          0.1       2.857143 2
[5]  {Station_entrance_and_service_desk} => {My_employment_number}     0.1     1          0.1       2.222222 2
[6]  {I_belong_to_area}                => {My_first_and_last_name}     0.1     1          0.1       2.857143 2
[7]  {I_belong_to_area}                => {My_employment_number}       0.1     1          0.1       2.222222 2
[8]  {Start_time}                      => {End_time}                   0.1     1          0.1      10.000000 2
[9]  {End_time}                        => {Start_time}                 0.1     1          0.1      10.000000 2
[10] {My_goals}                        => {Employee_name}              0.1     1          0.1      10.000000 2
```

Figure 8 Association rules

On the basis of these 10 rules it can be said that the fields in Figure 8 always occur together:

- Goal; Year
- My immediate boss; Year
- My immediate boss; Individual goal
- Station entrance and service desk; My first and last name
- Station entrance and service desk; My employment number
- I belong to area; My first and last name
- Start time; End time
- End time; Start time
- My goals; Employee name

Coverage measures the probability that a rule applies to a randomly selected transaction. It is estimated by the proportion of transactions that contain the antecedent (LHS) of the rule [50]. These 10 rules here can be applied to 10% of randomly selected digital forms.

Lift measure shows how many times more often X and Y occur together than would be expected if they were statistically independent. So in essence, it indicates how significant is the consequent with respect to the antecedent [50]. Rules are sorted by lift in Figure 9.

```
> inspect(head(sort(rule1,by="lift"),10))
     lhs                          rhs                       support confidence coverage lift count
[1]  {Start_time}              => {End_time}                0.1     1          0.1      10   2
[2]  {End_time}                => {Start_time}              0.1     1          0.1      10   2
[3]  {My_goals}                => {Employee_name}           0.1     1          0.1      10   2
[4]  {Employee_name}           => {My_goals}                0.1     1          0.1      10   2
[5]  {My_goals}                => {Customer_service_KPIs}   0.1     1          0.1      10   2
[6]  {Customer_service_KPIs}   => {My_goals}                0.1     1          0.1      10   2
[7]  {Employee_name}           => {Customer_service_KPIs}   0.1     1          0.1      10   2
[8]  {Customer_service_KPIs}   => {Employee_name}           0.1     1          0.1      10   2
[9]  {Carriage_type}           => {Belongs_to_projects}     0.1     1          0.1      10   2
[10] {Belongs_to_projects}     => {Carriage_type}           0.1     1          0.1      10   2
```

Figure 9 Association rules sorted by lift

It can be seen in Figure 9 that the likelihood of these fields occurring together on digital form is 10 times higher than just the appearance of consequent:

- Start time; End time
- End time; Start time

- My goals; Employee name
- Employee name; My goals
- My goals; Customer service KPIs
- Customer service KPIs; My goals;
- Employee name; Customer service KPIs
- Customer service KPIs; Employee name
- Carriage type; Belongs to projects
- Belongs to projects; Carriage type

There was a very high number of rules, 109 415, and very clear connection between two fields always occurring together and because of that redundant rules were removed. A rule is redundant if it is a subset of larger rules. It means that there are more general rules with the same or higher confidence and with same RHS, but at least one item removed from LHS [51].

After removing redundant rules, 191 non-redundant rules remained and a grouped matrix was created for the visualization of these rules in Figure 10.

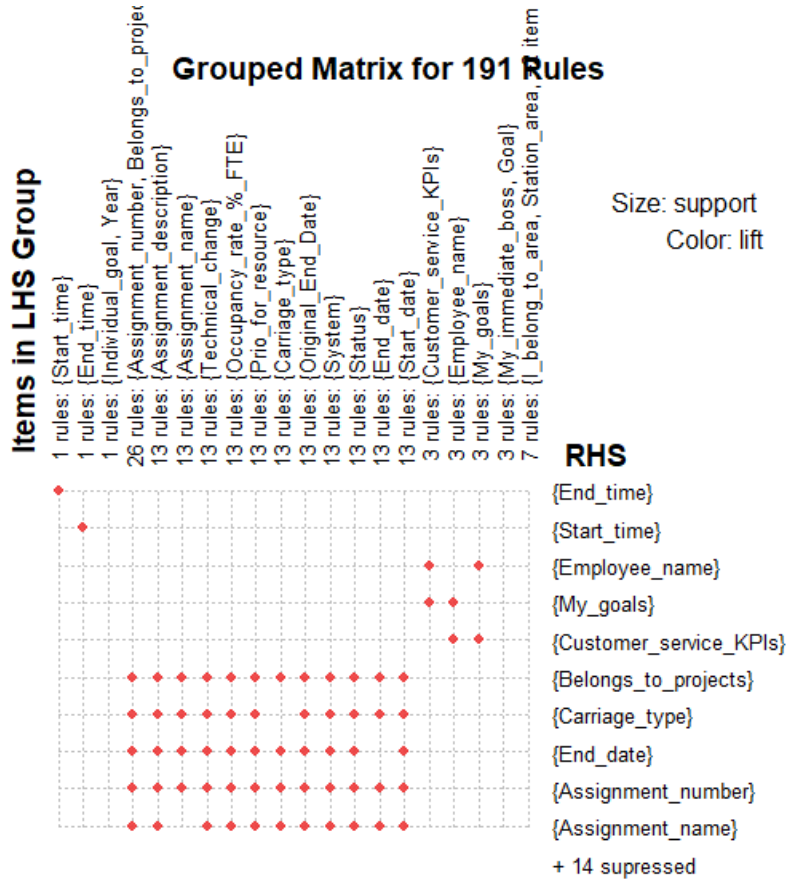

Figure 10 Grouped matrix for 191 non-redundant rules

It can be concluded from the size and colour of the support *"bubble"* that all fields have similar support and lift. The most ideal values are *Start time* and *End time*, which could also be seen in Figure 9, followed by *Customer Service KPIs* and *Employee name*, *My goals and Employee name*, *Customer Service KPIs* and *My goals, Employee name* and *My goals.*

4.3 Assigning security levels to non-redundant rules

The last step in the data mining process is presenting the analysis outcome in a format easily understandable for users, finding acceptable results and knowledge consolidation [36, p. 2].

Non-redundant rules discovered by data mining and Apriori algorithm were exported from RStudio into a .csv file. After that, non-redundant rules and metadata fields with previously assigned content categories and security levels were imported to PowerBI for

further data processing. Microsoft PowerBI is a business analytics tool for creating different reports and data visualization [52].

Data was pre-processed to a suitable format and a connection between tables with non-redundant rules and data security level was established via metadata field names.

After that, a report was created in PowerBI that includes the antecedent and consequent column from the non-redundant rules table, and data content category and data security level columns from the data security level table. A sample is presented in Table 8.

Antecedent	Consequent	Content	Data_Security_Level
Station_area	My_employment_number	station management	H
Station_area	My_first_and_last_name	station management	H
End_time	Start_time	train management	H
Start_time	End_time	train management	H
Station_area	My_employment_number	assets management	M
Station_area	My_first_and_last_name	assets management	M
Assignment_description	Assignment_name	employee management	M
Assignment_description	Assignment_number	employee management	M
Assignment_description	Belongs_to_projects	employee management	M
Assignment_description	Carriage_type	employee management	M

Table 8 Example of association rules sorted by security level

The purpose of the report was to simplify data security level evaluation through pre-evaluated metadata fields combinations. This report creates a data security evaluation list for Organization X, but it could also be extended and used by other companies with similar functions. The whole report is presented in Appendix 2 and in addition, relationships between metadata fields are presented on an interactive force-directed graph in Figure 11 Relationships between metadata fields. Visualization was done in PowerBI.

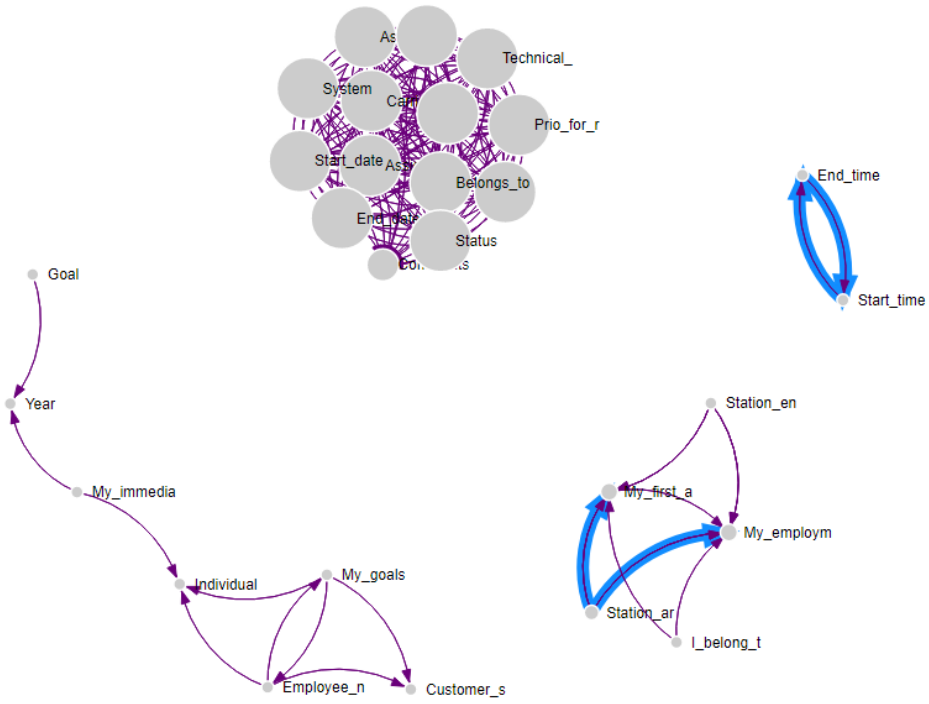

Figure 11 Relationships between metadata fields

In Figure 11 it can be seen that connected metadata fields converge into 4 groups. The connection was generated based on data category and security level weight. The weight of the connection shows the security level – thicker connection line means higher security level. There are 2 security levels in Appendix 2: high (H) and medium (M). Figure 11 presents the thickest connection between fields *Start time* and *End time*, *Station area* and *My first and last name*, *Station area* and *My employment numbe*r. It means that these connections have the highest security level (H) and all other connections have medium (M) security level. Same result can be seen in Appendix 2.

If user hovers the mouse over the connection, additional information about antecedent, consequent, data content category and security level can be seen. It is presented in Figure 12.

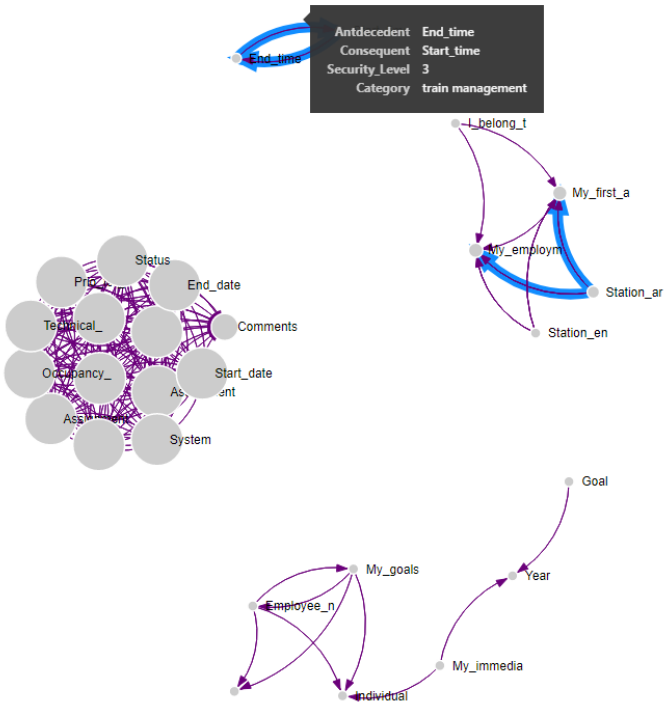

Figure 12 Additional information about field connection

Data security evaluation list can be used for evaluating data security by metadata field combinations on already existing or newly created digital forms. For example, as can be seen in Figure 11, if data content has been categorized as train management and the form includes fields Station area and My employment number, then the data security level is high (H) and this data can´t be published to cloud service according Table 5. But if the content of the same fields has been categorized as assets management, then data security level is medium (M) according to Table 5 and security measures described in chapter 3.2 must be implemented.

As mentioned earlier in the thesis, Organization X is planning to migrate from SharePoint Foundation 2013 to SharePoint Online, which enables the creation of sensitivity labels for data protection. Sensitivity labels help to classify and protect data according to business and compliance policies [53]. It is possible to use the data security evaluation list for defining and implementing sensitivity labels based on determined data content categories and security levels.

In addition, the data security levels evaluation list can be valuable input as pre-analysis for developing data security evaluation tool based on metadata fields.

5 Summary

An organization requesting to remain anonymous (Organization X) is planning a switch to cloud services, but it does not have internal data security regulations for setting the data security levels and measures.

The goal of the master thesis was to carry out an organization data security analysis based on information assets metadata fields. The main purpose of the thesis was to create the data security evaluation list based on metadata fields, which simplifies the decision-making process when publishing data to cloud services.

Data security analysis was performed in accordance with the three – level baseline security system ISKE (ISKE) [1]. First step was information asset inventory and as the result, all digital forms with their metadata fields were described. Metadata fields were divided into categories based on their content, security subclasses were evaluated, and security class value was formed. Security level was assessed based on security class. Limitations for cloud services were assessed and assigned to all metadata fields and digital forms. It was detected, which currently existing forms can be migrated to cloud service.

The problem was that all metadata fields must be evaluated one by one for achieving the information asset security level. Because of that the relationships between fields were explored for finding the rules, which could simplify the security level assignment. Data mining and association rules were used.

Apriori algorithm as one of the most influential technique in association rules field was used in RStudio and list of non-redundant connections between metadata fields was generated as a result. After that security levels were assigned to metadata field combinations by previously applied content categories.

The outcome is data security evaluation list based on metadata fields for Organisation X. It can be used for assessing and assigning security levels and measures for cloud services.

It could be used on already existing and newly created metadata fields and it can be valuable input as pre-analysis for developing data security evaluation tool based on metadata fields.

Metadata fields differ between organizations, but each organization can create their own data security evaluation list by using the same techniques for content analysis: data security level assessment by ISKE and association rules mining by Apriori algorithm.

References

[1] ‘Home - ISKE Portal - Version 8_06’. https://iske.ria.ee/8_06 (accessed Mar. 19, 2021).

[2] ‘Implementation manual for the THREE-LEVEL BASELINE SECURITY SYSTEM ISKE’. Accessed: Apr. 15, 2021. [Online]. Available: https://www.ria.ee/sites/default/files/content-editors/ISKE/iske-implementation-manual.pdf

[3] ‘https://www.microsoft.com/en-us/videoplayer/embed/RE1FGli?pid=ocpVideo0-innerdiv-oneplayer&jsapi=true&postJsllMsg=true&maskLevel=20&market=en-us’. https://www.microsoft.com/en-us/videoplayer/embed/RE1FGli?pid=ocpVideo0-innerdiv-oneplayer&jsapi=true&postJsllMsg=true&maskLevel=20&market=en-us (accessed Mar. 18, 2021).

[4] D. Zhe, W. Qinghong, S. Naizheng, and Z. Yuhan, ‘Study on Data Security Policy Based on Cloud Storage’, in *2017 ieee 3rd international conference on big data security on cloud (bigdatasecurity), ieee international conference on high performance and smart computing (hpsc), and ieee international conference on intelligent data and security (ids)*, May 2017, pp. 145–149. doi: 10.1109/BigDataSecurity.2017.12.

[5] R. V. Rao and K. Selvamani, ‘Data Security Challenges and Its Solutions in Cloud Computing’, *Procedia Comput. Sci.*, vol. 48, pp. 204–209, Jan. 2015, doi: 10.1016/j.procs.2015.04.171.

[6] S. Rajeswari and R. Kalaiselvi, ‘Survey of data and storage security in cloud computing’, in *2017 IEEE International Conference on Circuits and Systems (ICCS)*, Dec. 2017, pp. 76–81. doi: 10.1109/ICCS1.2017.8325966.

[7] J. Du, X. Zhang, H. Zhang, and L. Chen, ‘Research and improvement of Apriori algorithm’, in *2016 Sixth International Conference on Information Science and Technology (ICIST)*, May 2016, pp. 117–121. doi: 10.1109/ICIST.2016.7483396.

[8] ‘Juhend avalike pilveteenuste turvaliseks kasutamiseks avalikus sektoris’. Accessed: Mar. 19, 2021. [Online]. Available: https://www.ria.ee/sites/default/files/content-editors/ISKE/avalike-pilvede-kasutamise-juhend.pdf

[9] ‘RStudio | Open source & professional software for data science teams’. https://rstudio.com/ (accessed Apr. 09, 2021).

[10] M. G. Avram, ‘Advantages and Challenges of Adopting Cloud Computing from an Enterprise Perspective’, *Procedia Technol.*, vol. 12, pp. 529–534, Jan. 2014, doi: 10.1016/j.protcy.2013.12.525.

[11] H. Takabi, J. B. D. Joshi, and G. Ahn, ‘Security and Privacy Challenges in Cloud Computing Environments’, *IEEE Secur. Priv.*, vol. 8, no. 6, pp. 24–31, Nov. 2010, doi: 10.1109/MSP.2010.186.

[12] P. R. Kumar, P. H. Raj, and P. Jelciana, ‘Exploring Data Security Issues and Solutions in Cloud Computing’, *Procedia Comput. Sci.*, vol. 125, pp. 691–697, Jan. 2018, doi: 10.1016/j.procs.2017.12.089.

[13] F. Yahya, R. J. Walters, and G. B. Wills, ‘Protecting data in personal cloud storage with security classifications’, in *2015 Science and Information Conference (SAI)*, Jul. 2015, pp. 838–843. doi: 10.1109/SAI.2015.7237241.

[14] K. P. Singh, V. Rishiwal, and P. Kumar, ‘Classification of Data to Enhance Data Security in Cloud Computing’, in *2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU)*, Feb. 2018, pp. 1–5. doi: 10.1109/IoT-SIU.2018.8519934.

[15] R. Shaikh and M. Sasikumar, ‘Data Classification for Achieving Security in Cloud Computing’, *Procedia Comput. Sci.*, vol. 45, pp. 493–498, Jan. 2015, doi: 10.1016/j.procs.2015.03.087.

[16] X. Zhang, H. Du, J. Chen, Y. Lin, and L. Zeng, ‘Ensure Data Security in Cloud Storage’, in *2011 International Conference on Network Computing and Information Security*, May 2011, vol. 1, pp. 284–287. doi: 10.1109/NCIS.2011.64.

[17] K. A. Saed, N. Aziz, A. W. Ramadhani, and N. H. Hassan, ‘Data Governance Cloud Security Assessment at Data Center’, in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, Aug. 2018, pp. 1–4. doi: 10.1109/ICCOINS.2018.8510612.

[18] X. Chen, C. Chen, Y. Tao, and J. Hu, ‘A Cloud Security Assessment System Based on Classifying and Grading’, *IEEE Cloud Comput.*, vol. 2, no. 2, pp. 58–67, Mar. 2015, doi: 10.1109/MCC.2015.34.

[19] kaarins, ‘Introduction to SharePoint - SharePoint in Microsoft 365’. https://docs.microsoft.com/en-us/sharepoint/introduction (accessed Mar. 18, 2021).

[20] ‘ISO 23081-1:2017 Information and documentation. Records management processes. Metadata for records. Part 1: Principles’. Accessed: Mar. 19, 2021. [Online]. Available: https://www.evs.ee/Download/ViewBrowsingServiceSubscription?productId=113737&language=EnglishLanguage

[21] ‘Art. 4 GDPR – Definitions’, *General Data Protection Regulation (GDPR)*. https://gdpr-info.eu/art-4-gdpr/ (accessed Mar. 19, 2021).

[22] ‘Raamatupidamise seadus – Riigi Teataja’. https://www.riigiteataja.ee/akt/125052012016?leiaKehtiv (accessed Mar. 19, 2021).

[23] Organization X, ‘Document Management Manual’.

[24] ‘ISO/IEC 25010:2011 Systems and software engineering. Systems and software Quality Requirements and Evaluation (SQuaRE).System and software quality models’. https://www.evs.ee/Download/ViewBrowsingServiceSubscription?productId=30417&language=EnglishLanguage (accessed Mar. 19, 2021).

[25] ‘ISO/IEC 27001:2013 „Information technology. Security techniques. Information security management systems. Requirements“’, *EVS*. https://www.evs.ee/Download/ViewBrowsingServiceSubscription?productId=77970&language=EstonianLanguage (accessed Mar. 18, 2021).

[26] ‘ISO/IEC 27002:2017 “Information technology. Security techniques. Code of practice for information security controls“’, *EVS*. https://www.evs.ee/Download/ViewBrowsingServiceSubscription?productId=77974&language=EnglishLanguage (accessed Mar. 18, 2021).

[27] ‘COBIT | Control Objectives for Information Technologies’, *ISACA*. https://www.isaca.org/resources/cobit (accessed Mar. 18, 2021).

[28] ‘ITIL - ITIL’. https://www.itlibrary.org/ (accessed Mar. 18, 2021).

[29] ‘ISO/IEC 27017:2021 „Information technology. Security techniques. Code of practice for information security controls based on ISO/IEC 27002 for cloud services“’. https://www.evs.ee/Download/ViewBrowsingServiceSubscription?productId=135156&language=EnglishLanguage (accessed Mar. 19, 2021).

[30] ‘Bundesamt für Sicherheit in der Informationstechnik’, *Bundesamt für Sicherheit in der Informationstechnik*. https://www.bsi.bund.de/DE/Home/home_node.html (accessed Mar. 19, 2021).

[31] ‘INFOSÜSTEEMIDE KOLMEASTMELISE ETALONTURBE SÜSTEEMI ISKE Rakendusjuhend’. Accessed: Mar. 19, 2021. [Online]. Available: https://www.ria.ee/sites/default/files/content-editors/ISKE/iske_rakendusjuhend.pdf

[32] S. Samonas and D. Coss, ‘THE CIA STRIKES BACK: REDEFINING CONFIDENTIALITY, INTEGRITY AND AVAILABILITY IN SECURITY’, p. 25.

[33] ‘Infosüsteemide turvameetmete süsteem – Riigi Teataja’. https://www.riigiteataja.ee/akt/115092020015?dbNotReadOnly=true (accessed Mar. 21, 2021).

[34] M. Bouraoui, I. Bouzouita, and A. G. Touzi, ‘Hadoop based mining of distributed association rules from big data’, in *2017 18th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, Dec. 2017, pp. 185–190. doi: 10.1109/STA.2017.8314975.

[35] R. J. Roiger, *Data Mining: A Tutorial-Based Primer, Second Edition*. CRC Press, 2017.

[36] C. Song, ‘Research of association rule algorithm based on data mining’, in *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, Mar. 2016, pp. 1–4. doi: 10.1109/ICBDA.2016.7509789.

[37] L. Zhan, F. Yu, and H. Zhang, ‘A fast algorithm for mining temporal association rules based on a new definition’, in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Jul. 2017, pp. 1548–1553. doi: 10.1109/FSKD.2017.8392995.

[38] Q. Han, D. Lu, K. Zhang, H. Song, and H. Zhang, ‘Secure Mining of Association Rules in Distributed Datasets’, *IEEE Access*, vol. 7, pp. 155325–155334, 2019, doi: 10.1109/ACCESS.2019.2948033.

[39] S. Sriramoju, ‘OPPORTUNITIES AND SECURITY IMPLICATIONS OF BIG DATA MINING’, *Int. J. Res. Sci. Eng.*, vol. 3, pp. 44–58, Nov. 2017.

[40] F. Salo, M. Injadat, A. B. Nassif, A. Shami, and A. Essex, ‘Data Mining Techniques in Intrusion Detection Systems: A Systematic Literature Review’, *IEEE Access*, vol. 6, pp. 56046–56058, 2018, doi: 10.1109/ACCESS.2018.2872784.

[41] R. Mendes and J. P. Vilela, ‘Privacy-Preserving Data Mining: Methods, Metrics, and Applications’, *IEEE Access*, vol. 5, pp. 10562–10582, 2017, doi: 10.1109/ACCESS.2017.2706947.

[42] B. R. Mistry and A. Desai, ‘Privacy preserving heuristic approach for association rule mining in distributed database’, in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Mar. 2015, pp. 1–7. doi: 10.1109/ICIIECS.2015.7192972.

[43] M. Chaudhari and J. Varmora, ‘Advance privacy preserving in association rule mining’, in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Mar. 2016, pp. 2527–2530. doi: 10.1109/ICEEOT.2016.7755148.

[44] S. K. Solanki and J. T. Patel, ‘A Survey on Association Rule Mining’, in *2015 Fifth International Conference on Advanced Computing Communication Technologies*, Feb. 2015, pp. 212–216. doi: 10.1109/ACCT.2015.69.

[45] TopSharePoint.com, ‘SharePoint MythBusters: Top 5 Misconceptions of the Platform’. https://www.topsharepoint.com/sharepoint-mythbusters-top-5-misconceptions-of-the-platform (accessed Apr. 09, 2021).

[46] ‘Chapter 1 Introduction | R Packages’. https://r-pkgs.org/intro.html (accessed Apr. 09, 2021).

[47] M. Hahsler, B. Grün, and K. Hornik, ‘arules --- Mining Association Rules and Frequent Itemsets with R’, *J. Stat. Softw.*, vol. 14, no. 15, 2005, doi: 10.18637/jss.v014.i15.

[48] E. Neuwirth, ‘RColorBrewer: ColorBrewer Palettes’, Dec. 07, 2014. https://CRAN.R-project.org/package=RColorBrewer (accessed Apr. 09, 2021).

[49] ‘Market Basket Analysis using R’, *DataCamp Community*, Aug. 21, 2018. https://www.datacamp.com/community/tutorials/market-basket-analysis-r (accessed Apr. 12, 2021).

[50] P. D. McNicholas, T. B. Murphy, and M. O’Regan, ‘interestMeasure: Calculate Additional Interest Measures’, *Computational Statistics & Data Analysis*, Jun. 2008. https://linkinghub.elsevier.com/retrieve/pii/S0167947308001709 (accessed Apr. 09, 2021).

[51] ‘is.redundant function - RDocumentation’. https://www.rdocumentation.org/packages/arules/versions/1.6-7/topics/is.redundant (accessed Apr. 14, 2021).

[52] ‘What is Power BI | Microsoft Power BI’. https://powerbi.microsoft.com/en-us/what-is-power-bi/ (accessed Apr. 14, 2021).

[53] cabailey, ‘Learn about sensitivity labels - Microsoft 365 Compliance’. https://docs.microsoft.com/en-us/microsoft-365/compliance/sensitivity-labels (accessed Apr. 14, 2021).

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Külli Kool

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Data security analysis for cloud service in Organization X” supervised by Jaak Tepandi
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

10.05.2021

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – Data security evaluation list for cloud service

Antecedent	Consequent	Content_Category	Data_ Security _Level
Station_area	My_employment_number	station management	H
Station_area	My_first_and_last_name	station management	H
End_time	Start_time	train management	H
Start_time	End_time	train management	H
Station_area	My_employment_number	assets management	M
Station_area	My_first_and_last_name	assets management	M
Assignment_description	Assignment_name	employee management	M
Assignment_description	Assignment_number	employee management	M
Assignment_description	Belongs_to_projects	employee management	M
Assignment_description	Carriage_type	employee management	M
Assignment_description	Comments	employee management	M
Assignment_description	End_date	employee management	M
Assignment_description	Occupancy_rate_%_FTE	employee management	M
Assignment_description	Original_End_Date	employee management	M
Assignment_description	Prio_for_resource	employee management	M
Assignment_description	Start_date	employee management	M
Assignment_description	Status	employee management	M
Assignment_description	System	employee management	M
Assignment_description	Technical_change	employee management	M
Assignment_name	Assignment_description	employee management	M
Assignment_name	Assignment_number	employee management	M
Assignment_name	Belongs_to_projects	employee management	M
Assignment_name	Carriage_type	employee management	M
Assignment_name	Comments	employee management	M
Assignment_name	End_date	employee management	M
Assignment_name	Occupancy_rate_%_FTE	employee management	M
Assignment_name	Original_End_Date	employee management	M
Assignment_name	Prio_for_resource	employee management	M
Assignment_name	Start_date	employee management	M
Assignment_name	Status	employee management	M
Assignment_name	System	employee management	M
Assignment_name	Technical_change	employee management	M
Assignment_number	Assignment_description	employee management	M
Assignment_number	Assignment_name	employee management	M
Assignment_number	Belongs_to_projects	employee management	M

Assignment_number	Carriage_type	employee management	M
Assignment_number	Comments	employee management	M
Assignment_number	End_date	employee management	M
Assignment_number	Occupancy_rate_%_FTE	employee management	M
Assignment_number	Original_End_Date	employee management	M
Assignment_number	Prio_for_resource	employee management	M
Assignment_number	Start_date	employee management	M
Assignment_number	Status	employee management	M
Assignment_number	System	employee management	M
Assignment_number	Technical_change	employee management	M
Belongs_to_projects	Assignment_description	employee management	M
Belongs_to_projects	Assignment_name	employee management	M
Belongs_to_projects	Assignment_number	employee management	M
Belongs_to_projects	Carriage_type	employee management	M
Belongs_to_projects	Comments	employee management	M
Belongs_to_projects	End_date	employee management	M
Belongs_to_projects	Occupancy_rate_%_FTE	employee management	M
Belongs_to_projects	Original_End_Date	employee management	M
Belongs_to_projects	Prio_for_resource	employee management	M
Belongs_to_projects	Start_date	employee management	M
Belongs_to_projects	Status	employee management	M
Belongs_to_projects	System	employee management	M
Belongs_to_projects	Technical_change	employee management	M
Carriage_type	Assignment_description	employee management	M
Carriage_type	Assignment_name	employee management	M
Carriage_type	Assignment_number	employee management	M
Carriage_type	Belongs_to_projects	employee management	M
Carriage_type	Comments	employee management	M
Carriage_type	End_date	employee management	M
Carriage_type	Occupancy_rate_%_FTE	employee management	M
Carriage_type	Original_End_Date	employee management	M
Carriage_type	Prio_for_resource	employee management	M
Carriage_type	Start_date	employee management	M
Carriage_type	Status	employee management	M
Carriage_type	System	employee management	M
Carriage_type	Technical_change	employee management	M
End_date	Assignment_description	employee management	M
End_date	Assignment_name	employee management	M
End_date	Assignment_number	employee management	M
End_date	Belongs_to_projects	employee management	M
End_date	Carriage_type	employee management	M
End_date	Comments	employee management	M
End_date	Occupancy_rate_%_FTE	employee management	M
End_date	Original_End_Date	employee management	M

End_date	Prio_for_resource	employee management	M
End_date	Start_date	employee management	M
End_date	Status	employee management	M
End_date	System	employee management	M
End_date	Technical_change	employee management	M
End_time	Start_time	employee management	M
Goal	Year	employee management	M
My_goals	Customer_service_KPIs	employee management	M
My_goals	Employee_name	employee management	M
My_goals	Individual_goal	employee management	M
Occupancy_rate_%_FTE	Assignment_description	employee management	M
Occupancy_rate_%_FTE	Assignment_name	employee management	M
Occupancy_rate_%_FTE	Assignment_number	employee management	M
Occupancy_rate_%_FTE	Belongs_to_projects	employee management	M
Occupancy_rate_%_FTE	Carriage_type	employee management	M
Occupancy_rate_%_FTE	Comments	employee management	M
Occupancy_rate_%_FTE	End_date	employee management	M
Occupancy_rate_%_FTE	Original_End_Date	employee management	M
Occupancy_rate_%_FTE	Prio_for_resource	employee management	M
Occupancy_rate_%_FTE	Start_date	employee management	M
Occupancy_rate_%_FTE	Status	employee management	M
Occupancy_rate_%_FTE	System	employee management	M
Occupancy_rate_%_FTE	Technical_change	employee management	M
Original_End_Date	Assignment_description	employee management	M
Original_End_Date	Assignment_name	employee management	M
Original_End_Date	Assignment_number	employee management	M
Original_End_Date	Belongs_to_projects	employee management	M
Original_End_Date	Carriage_type	employee management	M
Original_End_Date	Comments	employee management	M
Original_End_Date	End_date	employee management	M
Original_End_Date	Occupancy_rate_%_FTE	employee management	M
Original_End_Date	Prio_for_resource	employee management	M
Original_End_Date	Start_date	employee management	M
Original_End_Date	Status	employee management	M
Original_End_Date	System	employee management	M
Original_End_Date	Technical_change	employee management	M
Prio_for_resource	Assignment_description	employee management	M
Prio_for_resource	Assignment_name	employee management	M
Prio_for_resource	Assignment_number	employee management	M
Prio_for_resource	Belongs_to_projects	employee management	M
Prio_for_resource	Carriage_type	employee management	M
Prio_for_resource	Comments	employee management	M
Prio_for_resource	End_date	employee management	M
Prio_for_resource	Occupancy_rate_%_FTE	employee management	M

Prio_for_resource	Original_End_Date	employee management	M
Prio_for_resource	Start_date	employee management	M
Prio_for_resource	Status	employee management	M
Prio_for_resource	System	employee management	M
Prio_for_resource	Technical_change	employee management	M
Start_date	Assignment_description	employee management	M
Start_date	Assignment_name	employee management	M
Start_date	Assignment_number	employee management	M
Start_date	Belongs_to_projects	employee management	M
Start_date	Carriage_type	employee management	M
Start_date	Comments	employee management	M
Start_date	End_date	employee management	M
Start_date	Occupancy_rate_%_FTE	employee management	M
Start_date	Original_End_Date	employee management	M
Start_date	Prio_for_resource	employee management	M
Start_date	Status	employee management	M
Start_date	System	employee management	M
Start_date	Technical_change	employee management	M
Start_time	End_time	employee management	M
Station_entrance_and_service_desk	My_employment_number	employee management	M
Station_entrance_and_service_desk	My_first_and_last_name	employee management	M
Status	Assignment_description	employee management	M
Status	Assignment_name	employee management	M
Status	Assignment_number	employee management	M
Status	Belongs_to_projects	employee management	M
Status	Carriage_type	employee management	M
Status	Comments	employee management	M
Status	End_date	employee management	M
Status	Occupancy_rate_%_FTE	employee management	M
Status	Original_End_Date	employee management	M
Status	Prio_for_resource	employee management	M
Status	Start_date	employee management	M
Status	System	employee management	M
Status	Technical_change	employee management	M
System	Assignment_description	employee management	M
System	Assignment_name	employee management	M
System	Assignment_number	employee management	M
System	Belongs_to_projects	employee management	M
System	Carriage_type	employee management	M
System	Comments	employee management	M
System	End_date	employee management	M
System	Occupancy_rate_%_FTE	employee management	M
System	Original_End_Date	employee management	M
System	Prio_for_resource	employee management	M

System	Start_date	employee management	M
System	Status	employee management	M
System	Technical_change	employee management	M
Technical_change	Assignment_description	employee management	M
Technical_change	Assignment_name	employee management	M
Technical_change	Assignment_number	employee management	M
Technical_change	Belongs_to_projects	employee management	M
Technical_change	Carriage_type	employee management	M
Technical_change	Comments	employee management	M
Technical_change	End_date	employee management	M
Technical_change	Occupancy_rate_%_FTE	employee management	M
Technical_change	Original_End_Date	employee management	M
Technical_change	Prio_for_resource	employee management	M
Technical_change	Start_date	employee management	M
Technical_change	Status	employee management	M
Technical_change	System	employee management	M
Station_area	My_employment_number	financial data	M
Station_area	My_first_and_last_name	financial data	M
Station_entrance_and_service_desk	My_employment_number	financial data	M
Station_entrance_and_service_desk	My_first_and_last_name	financial data	M
Employee_name	Customer_service_KPIs	personal data	M
Employee_name	Individual_goal	personal data	M
Employee_name	My_goals	personal data	M
I_belong_to_area	My_employment_number	personal data	M
I_belong_to_area	My_first_and_last_name	personal data	M
My_first_and_last_name	My_employment_number	personal data	M
My_immediate_boss	Individual_goal	personal data	M
My_immediate_boss	Year	personal data	M