Creating a Framework for GDPR-compliant Data Warehouse Using Data Protection by Design

How application of Data Protection by Design and Privacy by Design philosophies in Information Systems Management can help build GDPR-compliant systems?

Nataliya Godunok

A dissertation submitted to University of Dublin

in partial fulfilment of the requirements for the degree of

MSc in Management of Information Systems

2019

Abstract

Data is one of the most valuable assets in modern enterprises as it provides businesses with strategic information and supports the decision-making process. Nonetheless, management of organizations often has limited knowledge of types and volumes of data that they collect, where it is stored, how it is used, who can access it and who is responsible for the management of data strategy. Enforcement of the General Data Protection Regulation in the European Union has increased the level of responsibility that companies hold in regard to protection of personal data of users and has grown awareness of their privacy rights amongst individuals. This motivated numerous discussions between academics and industry professionals regarding applicability of the GDPR to information systems domains and feasibility of the proposed privacy frameworks and design techniques. Data analytics is one of the concerned areas affected by the Regulation, and companies aim to implement appropriate privacy strategies while minimizing risks and disruption to the business. Data Warehouses are strategic tools that are used in data analytics as they support business decisions by supplying answers for relevant questions. Therefore, it is important to investigate how such systems can be implemented with privacy in mind while still being able to fulfill its business function. Using design science research methodology, this study reviews available privacy frameworks, analyzes their compatibility with data protection principles and tests the application of relevant privacy techniques and strategies by evaluating the process of designing and architecting a Data Warehouse system. The proposed solution answers the question of how application of Data Protection by Design and Privacy by Design philosophies in Information Systems Management can help organizations build GDPR-compliant systems.

Acknowledgements

I would like to thank my supervisor Dr. Hugh Gibbons for all the assistance he provided throughout my research, for guiding to the right direction and being always available to answer questions when support was needed. I am also grateful to my classmates and lecturers of the course who were inspiring me on this journey for the last two years.

I also want to express my gratitude to my colleagues and managers at work who were supporting me in achieving my goals, encouraging to push the boundaries and motivating in times when it was hard to maintain the balance.

I am thankful to my husband Artem Krylov for proofreading sections of the dissertation, but more importantly for being patient, supportive and understanding throughout this long journey, for giving me time to spend on this study and especially for believing in me.

Declaration

I declare that the work described in this dissertation is, except where otherwise stated, entirely my own work, and has not been submitted as an exercise for a degree at this or any other university. I further declare that this research has been carried out in full compliance with the ethical research requirements of the School of Computer Science and Statistics.

Signed: Nataliya Godunok

Nataliya Godunok

6th May 2019

Permission to Lend and/or Copy

I agree that the School of Computer Science and Statistics, Trinity College may lend or copy this dissertation upon request.

Signed: Nataliya Godunok

Nataliya Godunok

6th May 2019

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BA – Business Analytics

BI – Business Intelligence

CEO – Chief Executive Officer

DPbD – Data Protection by Design

DPD – Data Protection Directive

DPP – Data Protection Principle

DSS – Decision Support Systems

DW – Data Warehouse

EC – European Commission

ENISA - European Union Agency for Network and Information Security

ETL – Extract, Transform and Load

EU – European Union

FIPs – Fair Information Principles

GDPR – General Data Protection Regulation

IAM – Identity and Access Management

ICO – Information Commissioner's Office

IS – Information System

IT – Information Technology

PbD – Privacy by Design

PETs – Privacy Enhancing Technologies

PII – Personally Identifying Information

PK – Primary Key

SDLC – Software Development Life Cycle

UK – United Kingdom

US – Unites States [of America]

1 Introduction

This dissertation is a product of extensive interdisciplinary research, where policy is coupled with engineering to propose credible techniques, processes and methods of planning, setting requirements, designing, architecting and building information systems compliant with relevant regulations, which in the context of this research is General Data Protection Regulation (GDPR). The GDPR is considered to be the most important change in data privacy regulations in the last few decades (*EU GDPR.ORG, n.d.*) as discussed further in greater detail in *Background and History* section of this chapter. Taking this into account, the study covers a broad range of topics and touches questions from multiple areas that are important to be discussed within the scope of this research in order to understand the problem and its relevance for the defined audience. *Appendix 1* depicts a mind map showing relations between the concepts and the issues discussed in this research while pointing out the questions worth answering in the context of each subject.

Apart from the GDPR, which comes from the policy and regulations side of the investigation, another major area of this study is the process of planning and architecting a Data Warehouse (DW) system in an enterprise that in turn represents the engineering and design component of research. This paper proposes to combine relevant frameworks and best practices from both areas, and thus contributes to the argument on how policy and engineering might coexist.

The first chapter of this dissertation introduces the discussion and defines the context of the research. It also provides background information on the research topic and describes the nature of the problem. It then goes on with explaining how the research question was developed, what are the secondary questions worth answering in the course of this study and what are the objectives and expected outcomes. It is important to define the audience that may benefit from this research, so the relevance and value of this study is also discussed in this chapter. Finally, it is recognized what is and what is not in the scope of this research, and then the chapter ends with the description of dissertation structure.

1.1. Background and History

Safeguarding personal data of European citizens is promoted and prioritized by European Commission (EC) with the same level of importance as other fundamental rights that are based on the values of equality, non-discrimination, inclusion, human dignity and democracy (*EC, n.d.(a)*). Businesses should be aware of their obligations in regard to the protection of personal data of their customers, employees and other stakeholders, and thus take appropriate actions to implement valid mechanisms, both technical and organizational (*Intersoft Consulting, n.d.*), to support these obligations since the GDPR (officially known as

Regulation (EU) 2016/679) was finally enforced by European Union (EU) Parliament in May 2018 following a two-year period that was given to organizations to prepare for the change.

The GDPR applies to any organization that processes personal data, whether automatically or manually, and even if the data in question is processed for another company. As stated by the EC, the GDPR applies if: "*your company processes personal data and is based in the EU, regardless of where the actual data processing takes place; or your company is established outside the EU but offers goods or services to, or monitors the behaviour of, individuals within the EU*" (*EU, 2018*). It may be concluded that the GDPR affects companies operating in the EU, regardless of their location; it regulates how businesses collect, store and process Personally Identifying Information (referred as PII data) of individuals.

While businesses must commit to implementation of new regulations, individuals (in the text of the GDPR officially referred to as "natural persons"), on the other hand, have now more control over their personal data and how it is shared and used. Amongst the obligations and rights that the regulation presents, EC distinguishes the following principles (*EC, n.d.(b)*):

- *Clear language*. Businesses should use clear and straightforward language and terms in their privacy policies.
- *Consent from user*. The user must clearly consent to the processing of their data, and businesses should not assume that silence means consent.
- *Transparency*. Users must be clearly informed if their data is transferred outside EU; businesses should distinctly define the purpose of data collection and processing; companies must inform users whether a decision based on their personal data was automated, as users should have possibility to contest it.
- *Strong rights*. Users should be informed in case of any data breaches within a set period of time; it should not be difficult for users to get access to the data that company possesses about them; if requested, users should either be able to move their data to another competing service or request their data deletion (also known as "the right to be forgotten").
- *Strong enforcement.* In case of non-compliance, businesses may be fined for up to 20 million EUR or 4% of a company's worldwide turnover.

The regulation replaces former Data Protection Directive (DPD), officially known as *Directive 95/46/EC*, that was originally put into effect by the EC in 1995 and envisioned same principles, with the difference that the latter one is now enforced by law. Even though the regulation may seem to be restrictive and put organizational and financial burden on small and medium size companies, the scandal around data of 87 million Facebook users that had been exposed to Cambridge Analytica in 2018 (*Lapowsky, 2018*) convinced the public that the GDPR is the right path to go on the way of protecting individuals' privacy.

From this perspective, one area worth exploring is how this regulation is applicable to Big Data and Data Analytics, and this research is focused on the Data Warehouse (DW) as a combination of tools, techniques and processes that organizations use to collect, store and process data.

1.2. Context of Research

The GDPR, its principles and impact have become a big topic of discussion that brings attention of researchers across multiple industries. Various aspects of the regulation are being looked at from different perspectives. During the month when the GDPR became enforceable by law, it was looked up more often in Google search than some of the most popular American celebrities Beyoncé and Kim Kardashian:

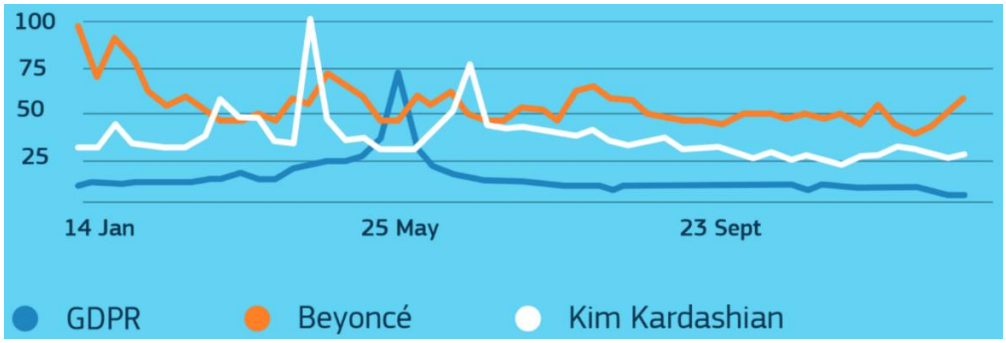


Figure 1: Interest rated between 0-100, based on number of searches on Google. Source: Google trends (EC, 2019; the GDPR Infographics)

Considering the background of the topic, it was decided that several aspects of the regulation are worth discussing in the context of academic research. This study in particular is based on Article 25 "Data protection by design and by default" of the GDPR (full text provided in *Appendix 2*), where it is investigated whether Data Protection by Design (DPbD) and Privacy by Design (PbD) principles can be applied in the domain of information systems. To narrow down the research question, it was decided that only a specific piece of technology must be explored, so it would fit into the scope and timeline of the study. As proposed in the previous section of this chapter, the research will focus on the implementation process of DW as a system that is widely used in conjunction with Business Analytics (BA) / Business Intelligence (BI) applications. Since organizations use various techniques and tools for collecting, storing, processing and analyzing data, and considering that the GDPR is now regulating all aspects of such activities, this study will explore how protection of individuals' data can be ensured at every stage of data management process. In this work, relevant literature regarding applicability of DPbD and PbD principles in system design, architecture and development is critically reviewed, existing frameworks analyzed, and reasonable conclusions are drawn in order to support the argument and provide a justifiable solution how the GDPR-compliant DW should be build.

1.3. Problem Description and Nature

Data nowadays is often referred to as "new oil", and predictions are that the amount of data generated globally will hit 44 zettabytes by 2020, which is ten times more than 2013's 4.4 zettabytes (*Kugler, 2018*). Access to accurate, valid and timely information may allow organizations to gain significant advantage in competitive race. Data is one of the most valuable assets of enterprises, and it helps companies make strategic business decisions. Until the GDPR was enforced, organizations could collect any amounts of data "just in case" and then use and reuse information available to them for various purposes whereas individuals might have not been fully aware of who holds which information about them, how this data is used and whether it is shared with any third-parties. It may be easy to assume that the more data one has, the greater the value it will generate. However, the GDPR now obliges organizations to think of data protection and privacy of their customers and other stakeholders at every stage of data management process, – this is now enforced by law. One point of discussion amongst researchers (discussed further in greater detail in *Literature Review* chapter) is that the GDPR tells organizations what must be done; however, it does not directly supply techniques or provide clear instructions on how exactly the Regulation must be implemented, thereby creating a ground for research in this area.

Companies are building their strategies and making important business decisions based on data available to them. Inaccurate or incomplete data may be misleading and cause significant disruption to the business, both financial and operational, so organizations should adapt to changes enforced by regulations and revise their data management and governance strategies (*Turner and Burbank, 2016*). To make right strategic decisions, management of organizations must analyze significant amount of data points, for which various BA/BI tools are used. The source of data for these tools are often Data Warehouses. The purpose of a DW is to consolidate data from multiple sources and allow business to make forecasts based on historical data and trends, thus a DW is one of the few tools that directly impact business decisions (*Poniah, 2010, p.15*), and therefore organizations must ensure that it is designed, developed and managed in line with the GDPR. Even though it may be argued that the text of the GDPR itself does not produce sufficient information on the implementation of the regulation in organizations, the Information Commissioner's Office (*ICO, n.d.(a)*), UK's independent authority established to uphold information rights, provides a comprehensive list of items and best practices that companies need to consider and which questions to answer when implementing appropriate organizational measures to ensure protection of data. Following recommendations of ICO and having reviewed works of other researches in areas of DPbD and PbD, this study aims to discover applicable privacy framework that can be projected on the process of DW design and development.

1.4. Research Questions and Objectives

To mitigate uncertainty around impact of the GDPR on business, companies must implement appropriate organizational and technical measures to ensure data protection (*Intersoft Consulting, n.d.*). The breadth of the research topic, as illustrated in *Appendix 1*, leads to the following secondary areas and related sub questions that need to be explored:

- *The GDPR and Data Protection by Design and by Default*:
  - What is PII data, what are the examples and why it requires protection?
  - What are the principles of DPbD and PbD?
  - Why privacy is important, who benefits from it, and can it be hardcoded?
- *Data as an Asset*:
  - How data contributes to the decision-making process?
  - Why having accurate data is important for business?
- *Data Analytics*:
  - Which data businesses collect, store, process and analyze, and how?
  - How above-mentioned activities may interfere with privacy?
  - Which tools, techniques and processes are used in data analytics?
- *Data Warehouse systems – what it is and its purpose*:
  - Which role DW plays in data analytics, and how DW benefits business?
  - What are the main DW architecture and design best practices?
  - What are major privacy and security concerns of DW?

Keeping in mind the above scope, primary research question was formulated for the study:

How can Data Protection by Design and Privacy by Design philosophies help organizations build GDPR-compliant Data Warehouses?

The final goal for this dissertation is to create a credible artefact, which would become a reference model for application of privacy design techniques in DW architecture. Below secondary objectives can form a list of activities that will contribute to the goal:

- Identify credible sources and review literature on the related topic
- Analyze available privacy design and data protection frameworks
- Chose suitable DW architecture that resonates with the GDPR principles
- Outline DW planning and design process using privacy frameworks

1.5. Relevance and Value of Research

This dissertation will be of value for multiple parties: academics, information systems, business and regulatory management. Extensive literature review resulted in a comprehensive list of credible resources, which students may use in their future research.

Analysis of existing privacy frameworks may be of use for management of organizations planning to implement relevant processes across multiple business units. Methods and techniques, discussed in this research could be of particular interest for project managers, Senior Management of IS/IT function in organization, business analysts, BI application designers/developers, technical architects and other individuals involved in the DW planning, development and maintenance activities.

While it is possible to find supporting literature about general concepts of privacy and data protection and how they coexist with technology, there are very few references related to application of privacy design techniques specifically in development of DW systems. Even though, due to limitations of this study, there will be plenty of areas for future studies, the primary research question contributes to the academic discussion with its uniqueness and distinctively defined scope.

1.6. Scope of Study

Overall, broad range of questions across several related topics are explored in this dissertation. While it is important to provide enough background, with such breadth of the study research question was narrowed down to a more specific problem. The GDPR covers many areas of data management process, but in the course of this research only *Article 25 'Data protection by design and by default'* (*Appendix 2*) and *Article 5 'Principles relating to processing of personal data'* (*Appendix 3*) are reviewed from the policy and design perspective. It is argued that both mentioned articles frame the core concepts of the GDPR, and implementation of their principles into systems development process will help companies become GDPR-compliant. From the engineering perspective, this research looks into planning, design and architecture of DW system and investigates application of privacy design techniques in this process. As DW systems directly contribute to the decision-making process by providing tools for the business to quickly access relevant sets of data, is it important to ensure that the processes used for collecting, storing and processing that data are compliant with relevant regulations.

As the breadth and depth of the research is defined, it is also worth noting what is not in scope of this study. Due to technical knowledge limitations, the study will not focus on technical details of recommended methods, but rather dive deep into organizational processes and design of overall framework for application of privacy design techniques.

1.7. Dissertation Timeframe and Chapter Roadmap

This research study was conducted between October 2018 and May 2019. During October and November, literature was reviewed on the topic of interest, and the major area of study was identified. After several meetings with supervisor, it was decided that research question

must be narrowed down, considering the timeframes of research and available resources. When in December more specific research topic was defined, extensive literature review was conducted on more relevant areas of study, while focusing on answering secondary questions and determining research value. During January and February, frameworks that constitute major part of the study were selected and relevant core architectures analyzed. In March, examples and processes for the theory testing part of dissertation were outlined, and finally all chapters drafted throughout April.

Below table summarizes the structure of this dissertation:

1 Introduction	Chapter I introduces the research topic, context of the study and general background. It also defines the problem and its nature, how the idea developed, why it is important and for whom. It clarifies what is in scope and out of scope of the study, outlines research question and objectives and provides details on the research timeframe.
2 Literature Review	Chapter II provides critical analysis of the related literature in the research field and covers major relevant theories, while positioning the research question in this context.
3 Research Methodology and Design	Chapter III outlines how the research questions can be answered, which frameworks are applied and why. It justifies the methods used and approaches chosen. The chapter ends with description of issues encountered during the process and limitations faced.
4 Framework Analysis and Application	Chapter IV aims to produce the final artefact set out in the objectives of the dissertation, using chosen frameworks and techniques. Then, analysis of the design theory is performed, and findings communicated.
5 Conclusions and Further Research	Chapter V concludes the dissertation – it reports the findings, interprets what was discovered, critically analyses the results and offers possible future directions for research in this area.
Appendix 1	Mind Map of Research Areas and Their Relations
Appendix 2	Article 25 of the GDPR: Data protection by design and by default
Appendix 3	Article 5 of the GDPR: Principles relating to processing of personal data
Appendix 4	List of questions that DW can answer as per Song and LeVan-Shultz (*1999*), categorized and tested against DPPs
Appendix 5	Base Star Schema for e-commerce sales as per Song and LeVan-Shultz (*1999*)

Table 1: Dissertation Chapters Roadmap

2 Literature Review

2.1 Introduction

Literature Review chapter of this dissertation outlines the main ideas and sets the base for further research, where the theory described in this dissertation can be experimentally tested. Researchers from multiple disciplines are discussing the phenomenon of privacy and ethics in engineering and systems design. Different perspectives for the topic open the discussion how policy and engineering may coexist. The purpose of this chapter is to review and analyze existing literature, get familiar with main concepts, major theories and concerns of researchers in the field, define applicable frameworks and set the ground for placing the problem discussed in this study in the context of the literature. In the course of this research numerous types of literature were examined and analyzed: books, journal and academic articles, blogs, online publications, extracts from related legislation and other resources.

As presented in *Appendix 1*, to understand the scope of the problem, to be able to dive deep into relations between the topics and then find the suitable solution, the research was divided into several parts, whence secondary questions were extracted. The following areas were explored as part of this study:

- The GDPR, its background and implications for both individuals and organizations, definitions and core principles, why it is important and to whom.
- PII data in the context of the GDPR, why it requires special protection, how organizations collect and use this data and what can be improved.
- The lifecycle of data in organizations, in which ways data is used, who decides its value and how it can help the business make strategic decisions.
- Common DW practices and techniques, how they raise GDPR-related concerns and how they may be addressed by applying privacy design strategies.
- Data Protection by Design and Privacy by Design principles, standards, design patterns, techniques and frameworks.
- Applicability of privacy and ethical engineering in modern organizations, privacy and security concerns in big data and analytics.

This chapter sets out the argument, compares the theories and shapes the importance of interdisciplinary research in management domain. One of the objectives of this chapter is to show how this study was developed and topics connected, and how the original idea for this research emerged and evolved. It also intends to respect the boundaries and limitations of research described in the *Introduction* chapter of this paper – omitting major technical details and focusing on the organizational side of the question. Though, certain technicalities will be involved, as without them the discussion would not be possible.

2.2 GDPR: Principles and Implications

It is worth noting that the GDPR is not the first and only regulation that dictates data protection standards. In its essence, it is a successor of *Directive 95/46/EC*, also known as The European Data Protection Directive, which was adopted in 1995. The aim of the Directive was to protect PII data and set the standards in privacy regulations, same as the GDPR. The major difference between the two is that the GDPR is now a legal requirement and non-compliance leads to fines, up to maximum of 20 million EUR or 4% of previous year annual global turnaround, whichever is higher (*Regulation (EU) 2016/679*). Countries in the EU must work on the development and implementation of the relevant legislation, and companies that work with personal data of EU citizens must ensure that their processes, tools as well as employees are compliant.

Considering that technology has progressed significantly over the last decades, one might assume that data breaches could occur more easily in the past. However, an interesting observation from this study was that it was hard to find information in the media about complaints and reports related to non-compliance with DPD. In the past few years, on the other hand, questions of privacy, protection of individuals' data and user trust have gained much more visibility despite the technological advancement and improvements in the field of automatic prevention of cyberthreats and protection against data breaches. Researchers from many areas are actively investigating the impact of the GDPR on different spheres of our lives nowadays and arguing regarding best practices and feasibility of its [GDPR] application. Few assumptions can be made about why there are more news about organizations being hacked and their data stolen. One is that companies are now required to report personal data breach (art.33 of GDPR) to the supervisory authority, while at the time of DPD this was not mandatory by law, and that have brought attention of the public and media to the subject. Another explanation is that an average reasonable person in developed society would have some doubts: if data is handled properly, if there is no bias and no reasons to be worried about the fact how data is used, then why implement such regulations? Level of trust to organizations that collect personal data of their users has dropped, and there are reasons to believe that this is due to information reported by media and news about cyberattacks, lost and stolen data. Hoare (*2018*) outlined in the article for *Irish Examiner* that in only seven months after the GDPR came into effect, more than 3,200 breaches of data have been reported to the Data Protection Commission, including a number of high-profile breaches, involving companies such as Twitter and Facebook. The Breach Level Index (*n.d.*) online portal provides historical information about data records lost or stolen since 2013, which is 14 billion records that equals to the frequency of 74 records per second. Amongst industries that had most data breaches are technology, social media, retail and government. It is worth noting that social media jumped to the top only in

the first half of 2018, which is most likely related to the major scandal around Facebook and Cambridge Analytica (*Lapowsky, 2018*). *Figure 2* represents the total number of records lost each month from 2013 until 2018, as per The Breach Level Index:

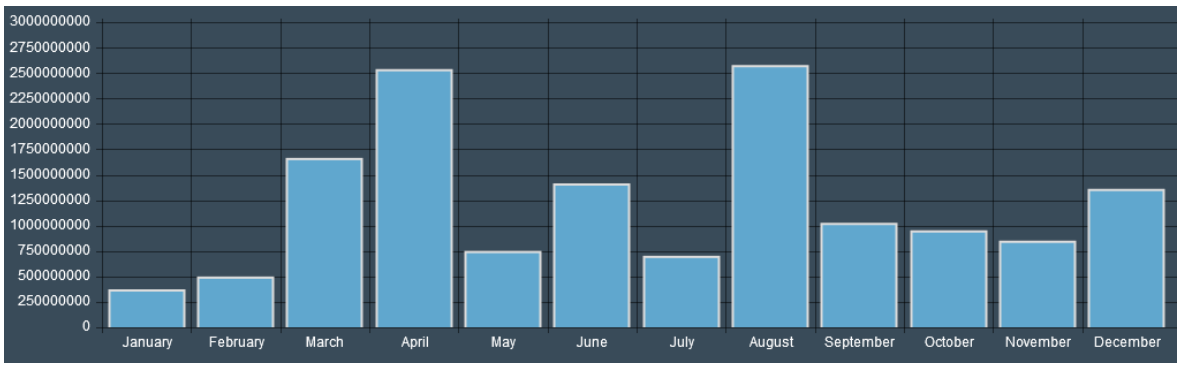


Figure 2: Total Records Lost by Month, 2013-2018 (The Breach Level Index)

Considering the above, the assumption regarding loss of user trust seems more than reasonable, and in that context, the GDPR might be the right way to move forward. Not only does the Regulation apply strict rules for organizations around the data collection, management and processing activities, but also provides assurance to individuals that they may now have more control of their personal data and how it is used (*Greengard, 2018*). Thus, the argument around how the Regulation should be implemented has begun.

An average reasonable person might think that enforcing the law might solve the issue of privacy and expect that prior to applying the Regulation, concerns from professionals around different industries had been reviewed and best practices on the implementation process had been agreed. This does not seem to be the case. Greengard (*2018*) notes that the impact and implementation recommendations of the GDPR is interpreted differently by professionals working in different areas and fulfilling different roles (legal, data analytics, system/software development), which raises a level of frustration amongst industry specialists: "*It is simply not possible to be 100% compliant*". It is argued that the Regulation puts companies in harsh conditions, as they must spend significant amount of time and resources to make sure they are compliant with the standards that are believed to be inconsistent with the way business is done online. The Economist (*2019a*) quotes a British group *Privacy International* that some organizations like banks may have valid reasons for collecting certain types of information that would be considered PII data for fraud prevention (for example, IP addresses and payment methods); however, companies involved in advertising cannot have "legitimate interest" in building their whole business model on gathered data. If such complaints are entertained by authorities, a large number of businesses could be at risk. At the same time, it was reported that few months before the GDPR came into effect, Facebook prompted its users to agree with the updated terms and conditions. Should customers not agree with the new terms, they would no longer have

access to their Facebook, Instagram and WhatsApp accounts (*The Ecomonist, 2019*), which was argued to be neither valid, nor appropriate option. In light of this dispute, McDougall (*2019*), Executive Director for Technology Policy and Innovation at the ICO of UK, offered a sound question that organizations must ask themselves: "*How much personal data, if any, is necessary for the system to function effectively?*". Considering the issues discussed above, it is assumed and suggested to test in this research that if a system is planned and designed with privacy in mind, then the above question would not apply.

To fully understand the principles discussed further in this research, some important concepts and terms used in the context of the GDPR must be explained. First of all, the main subject of this Regulation is protection of personal data of individuals, which is considered to be everyone's fundamental right (*Regulation (EU) 2016/679*). This type of data can be used whether on its own or with conjunction with other information to identify an individual, that is why it requires special protection – to ensure privacy and confidentiality as a fundamental right. Thus, anonymized or generalized data with higher level of granularity (the concept of granularity discussed further in the context of DW design and architecture) is not a subject to protection by the GDPR. Next section of this chapter is specifically dedicated to PII data and how to identify it.

Amongst other key definitions, special attention should be given to "controllers" and "processors", where the GDPR applies to both. Information Commissioner's Office (*n.d.(b)*) suggests a simple explanation of roles and responsibilities of both. These are strictly defined in the context of the Regulation. It can be concluded that the responsibilities of both roles of controller and processor are defined by a data management lifecycle, as each of them have a specific function. BSI (*n.d.*) define PII controller as the one who "*collects personal information and determines the purposes for which it is processed*", and PII processor – as someone who "*processes personal information on behalf of and only according to the instruction of the PII controller*". Data controller is not always one organization, and in cases where several companies act as controllers they are often known as co-controllers (this is where data-sharing agreements might be needed (*BSI, n.d.*)). It is also possible that the same company can be both controller and processor, and in the context of this research example of such scenario will be reviewed.

As for the main principles of the GDPR, various authors interpret Article 5 "Principles relating to processing of personal data" of the Regulation in a different way (full text of art.5 available in *Appendix 3*). In some sources, which were generally published years prior to the GDPR, these principles may be called by other terms even though they address same concepts of privacy: McCallister et al (*2010*) provide a list of eight "Fair Information Practices" (FIPs), Cho et al (*2015*) offer eleven "Privacy Principles", ENISA (*2018*) also talks about eleven

principles of privacy information security standard in the context of ISO/IEC 29100 privacy framework and Chessel (*2014*) suggest "ethical awareness framework". *Table 2* compares the sources that cover privacy principles and specifies how authors describe them as opposed to how this is presented in the GDPR. It is worth nothing that each individual article of the Regulation is only responsible for addressing a specific set of concerns, so they cannot be taken out of the context and investigated discretely. However, some of them frame a backbone of the GDPR, and it may be argued that removing certain articles would have caused a significant impact on the meaning of the directive as a whole.

	McCallister et al, 2010	*Cho et al, 2015*	*ENISA, 2019*	*Hoepman, 2014*	*Chessell, 2014*	*Chalcraft, 2018*	***GDPR***
Consent and choice		**x**	**x**	**x**	**x**	**x**	**x**
Purpose legitimacy and specification	**x**	**x**	**x**	**x**	**x**		**x**
Collection and use limitation	**x**	**x**	**x**		**x**	**x**	**x**
Data minimization		**x**	**x**	**x**			**x**
Use, retention and disclosure limitation		**x**	**x**	**x**	**x**	**x**	**x**
Accuracy and quality	**x**	**x**	**x**	**x**	**x**		**x**
Openness, transparency and notice	**x**	**x**	**x**	**x**	**x**	**x**	**x**
Individual participation and access	**x**	**x**	**x**	**x**	**x**	**x**	**x**
Accountability	**x**	**x**	**x**	**x**	**x**	**x**	**x**
Information security and protection	**x**	**x**	**x**	**x**			**x**
Privacy compliance		**x**	**x**	**x**		**x**	**x**
The right to be forgotten				**x**			**x**

Table 2: Comparison of privacy principles in the reviewed literature (summarized by author or this dissertation by aggregating reviewed literature)

Amongst researchers that investigate the issue of privacy and FIPs, Ann Cavoukian is recognized as a pioneer in attempting to build a reference privacy framework to be used for developing more specific criteria for application of privacy techniques in information systems. In 2009, Cavoukian posted her 7 Foundational Principles for Privacy by Design, and since then this framework received a lot of attention. Some scholars criticize either the framework or the concept of privacy in engineering for being too vague and generic concluding that it cannot be translated into technical requirements easily (*Koops and Leenes, 2014; Blix et al, 2017; Van Dijk et al, 2018*). Others take the opportunity to base their research on these principles, creatively explore the components of the framework in more detail and look at this topic from new perspectives – differentiating between PbD and

DPbD, distinguishing between 'privacy' and 'security', separating the concepts of 'privacy-by-policy' and 'privacy-by-architecture' or building privacy design patterns and strategies (*Spiekermann and Cranor, 2009; Kroener and Wright, 2014; Hafiz, 2016; Hoepman, 2018*). Works of the above cited authors as well as other resources are discussed later in this chapter looking at unique perspectives on the subject of translating privacy requirements into technical solutions.

2.3 PII Data: Privacy Matters

In GDPR as well as other sources PII data is also referred to as "personal data", which for the purposes of the Regulation means "any information relating to an identified or identifiable natural person ('data subject')" (*Regulation (EU) 2016/679, art.4*). Based on the examples provided in the Regulation, it is possible to compile a list of identifiers that can be labeled as PII data. For example:

- First name, middle names, surname, date of birth, phone number, address, email address, whether on their own (some names can be relatively rare) or in conjunction with one another can be used to identify an individual.
- An identification number: ID of passport, social card, driver license, student ID coupled with the college name, employee ID in conjunction with the company name or any other document/correspondence in relation to an individual.
- Location data. This can be either physical location identified via GPS navigation, or online identifiers of a user/individual (e.g., IP address, which can typically be used to physically locate the person, too).
- Financial information: bank account numbers, credit/debit cards.
- "Factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (*ibid, art.4*).

Some authors, whether prior or after GDPR publishing, state that it may be hard to define PII data, and some definitions can be vague. Narayanan and Shmatikov (*2010*) argue that email addresses or phone numbers do not fall under PII category. To think critically, at first sight it may depend on the context. Email address can be a distribution list that contains a group of individuals, so one cannot directly identify a person knowing that email only. Though on the other hand, the distribution list may direct the message to a specific group of people, for example, to a marketing team of the company. If it is possible to find out who is working in that marketing department, and, let us say, when receiving a response back from the team one can see a signature on the return email, then the possibility of identifying an individual is significantly increased, even though originally only the group email address was known. Another related example is that one can simply use their name or last name in

the local part[1] of the email address. Even though there is no guarantee that the person behind a specific email address has in fact the same name and last name as stated, GDPR suggests that if in doubt whether specific data is PII or not, then this piece of data should be treated as PII due to possible interpretation. Same applies to phone numbers. A phone number can be, for example, a reception number of a hotel, however one can find information about the employees who work at that reception, which automatically makes it easy to identify an individual who picks up the phone. In these examples, original item of data (phone number or email address) cannot be used to identify a unique person, but in conjunction with other knowledge, it will become PII data and require special handling.

Sweeney (*2002*) provides a fair example, how by combining data from two unrelated sources she identified the governor of Massachusetts at that time and his medical records. The leftmost circle in *Figure 3* represents information that The National Association of Health Data Organizations (NAHDO) suggested hospitals to collect since certain states had legitimate reasons to do so. As part of this, the Group Insurance Commission (GIC), who is responsible for buying health insurance for state employees, collected relevant information for about 135,000 individuals and their families. The data was believed to be anonymous, so GIC provided a copy to researchers and the industry. Then the author (*Sweeney, 2002*) was able to purchase the voter registration list for Cambridge Massachusetts, from which the data is presented in the rightmost circle in *Figure 3*. It was revealed that both sources of data could be linked with the ZIP code, birth date and gender. The governor's medical records were in GIC data, he lived in Cambridge Massachusetts, and according to the voter list, only six people had his particular date of birth, while three of them were men, and he was the only one with his 5-digit ZIP code.

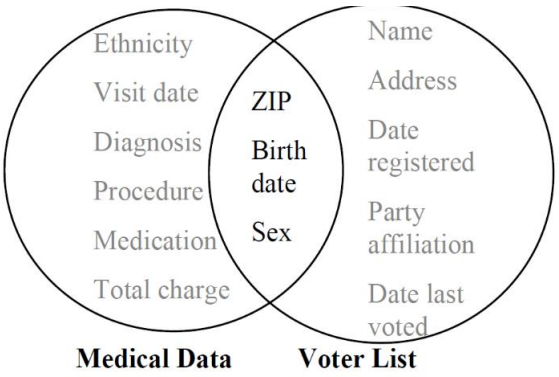


Figure 3: Linking unrelated sources to re-identify data (Sweeney, 2002)

As Sweeney (*2002*) states, the above example is a demonstration of re-identification by linking shared attributes and concludes in her research that the greater number of individual

[1] An email address such as John.Smith@example.com is made up of a local-part, an @ symbol and a case-insensitive domain (https://en.wikipedia.org/wiki/Email_address)

lines of data is available, the more anonymity that set of data provides. There are various privacy design techniques available as industry recommendations that allow to address many concerns regarding data handling practices and GDPR compliance. Data anonymization, as well as other privacy design techniques, is discussed later in this chapter.

Since it is possible to identify a person by combining data from datasets, which originally were believed to be anonymous and even were unrelated, the need to regulate activities related to data collection, sharing, processing and disclosure was identified and addressed by the GDPR. Potential data loss and data breaches not only may cause financial loses (where attackers who gained access to data may try to sell it back to the company, or even to their competitor), but also loss of trust, as discussed earlier in this chapter. Gemalto (*2017*) found from the survey, which was conducted on their behalf, that 67% out of 10,000 surveyed customers worldwide would not do business with the company that experienced data breach and 69% feel that companies do not take customers' data security seriously. Nonetheless, it appears that not only negligence with regard to relevant processes and implementation of security measures can harm business, but also, if attacker gets access to sensitive data, individual's privacy is affected, too. One of the most dangerous cyberthreats nowadays is identity theft (*Symantec, 2018*). With the growth of social networks, it has become easy to find enough information about individuals to impersonate them for the purpose of fraud. Even though organizations who collect data about individuals do have specific responsibilities regarding protection of this information, it is not enough just to rely on businesses in keeping the data secure. Gemalto (*2017*) reported that based on their survey, individuals are happy to make companies responsible, without making any effort to adequately secure their own data themselves. 56% still use the same password for multiple accounts, and while businesses succeeded in offering better level of security with Multi-Factor Authentication (MFA), 41% of respondents are not securing their accounts with this technology, leaving themselves vulnerable to potential breaches.

It may be concluded that to avoid data breaches, identity thefts and other cyberattacks, strategies and best practices for application of privacy and data protection techniques and principles must be diligently followed by both – companies and users. Both parties are interested in safe environment for systems and for data, thus collaboration is required.

2.4 Business Analytics and Privacy: Data as an Asset

Over the past decade, the statement "data is the new oil" has become increasingly popular (*Kugler, 2018*). Companies like advertising agencies build their whole business models on data, which is why they now face certain barriers in areas of their operation due to regulatory restrictions and privacy concerns. In fact, big data benefits all – from users to enterprises, from offering personalized experience to launching new opportunities in new markets after

analyzing behaviours and habits of consumers. Tene and Polonetsky (*2013*) discuss the use of big data for consumers and how analysis of large sets of information, which are not necessarily directly related to one another, can benefit areas such as healthcare, mobile and online business, smart grid, traffic management, retail and payments. Authors provide a strong example of a groundbreaking discovery that was made by a professor of medicine and bioengineering and his colleagues at Stanford University only using statistical analysis and data mining techniques to identify patterns in large techniques. In their paper, it has been noted that due to limited resources it is not always possible to perform every possible test on the medicine and check reaction for every potential interaction with other drugs. By utilizing datasets maintained by the Food and Drug Administration regarding drugs approved for use, the professor and his colleagues created a "symptomatic footprint" for diabetes-inducing drugs, searched for that footprint in interactions between pairs of drugs which were not known for causing such effect on their own, if taken alone, and discovered that four pairs of drugs were found to produce the footprint. Out of those, they decided to investigate Paxil and Pravachol more in details, since they were known to be the most commonly prescribed drugs. After approaching Microsoft Research with request to examine Bing search engine logs on the subject of word searches related to the "symptomatic footprint" together with both drugs names, they compared the results with searches of just the names of drugs without symptoms. Their research hypothesis was supported by the big data set received from the search engine. Users who searched names of both drugs together were much more likely to search for diabetes-related side effects than users who only searched for one of the drugs. As mentioned later in the paper (*Tene and Polonetsky, 2013*), this research was potentially life-saving for approximately one million of patients in the Unites States, who were prescribed both drugs.

Seeing the benefits of big data and data analytics, organizations are dedicating a fair amount of time and resources to attempt derive potential value of the data they collect and analyze (*Sidgman and Crompton, 2016; Chalcraft, 2018; Kugler, 2018*). While some researchers might still argue about the definition of data analytics, since the concept has evolved over time and now affects multiple industries, the overall understanding of the purpose of data analytics is that it helps organizations manage risks, maximize profitability, optimize resources, and identify and pursue new opportunities (*Chalcraft, 2018*). Such opportunities bring another question: what is the actual value of data? Machine learning and artificial intelligence significantly enhance the possibility of extracting value from structured and unstructured data, however the question how much it is worth remains open (*Kugler, 2018*). Sidgman and Crompton (*2016*) suggest that organizations should consider formal data valuation practice, which they believe will fix the root cause of the privacy failure and make it easier to manage, utilize and protect data. If companies knew how much their

data was worth and treated it with the same importance as other assets, they would also view privacy as a business issue, instead of just a compliance requirement; and with an execution of a formal data valuation, lawmakers, regulators and business leaders would have justified reasons to invest in data protection (*Sidgman and Crompton, 2016*). It is also argued that lack of formal valuation of data that organizations possess and its translation into financial presentation prevents the market from seeing the true value of each individual enterprise. Valuing data separately from systems used to collect it, equipment used to store it, or processes used to manage it throughout its life cycle, is not an easy task due to its intangible nature. The GDPR is changing the way how organizations collect their data and how they design and use their corporate databases; at the same time, digital technologies change the way how privacy is seen by consumers and how they perceive the importance of having the knowledge about who has access to their PII data, how it is collected an used – consumers are now more aware of the value of their personal data than before (*Greengard, 2018*). Kugler (*2018*) argues that customers are potentially willing to share their data if they get something in return; though, the benefits must be tangible (e.g., discounts) versus intangible such as product recommendations or supposedly simplified ordering.

Since data is a valuable asset (even though, intangible), a great responsibility falls on companies who use data that they collect in their benefit, as concerns regarding ethics and privacy arise – the GDPR sets a very specific set of requirements for businesses to be considered compliant, where purpose, collection and storage limitation, data minimization and other principles form a base of the Regulation. Industry leaders suggest that organizations must view their data strategically (*Van Hoof, 2017*). Sidgman and Crompton (*2016*), while talking about their proposal regarding implementation of data valuation process and its possible adoption by enterprises, suggest that strategic approach to data management would help organizations to be more cost efficient and generate new opportunities if risks are managed properly. Kugler (*2018*) insists that companies that work with data of EU citizens would need to reevaluate the scope of their data strategies or incorporate privacy-enhancing technologies (PETs) into their core business processes.

Even though, the GDPR answers the question “what needs to be done?”, it does not provide clear guidance on “how it needs to be done?”. There is a lot of research focused on the consumer side of the problem, providing insight into opinion of individuals regarding data handling practices, exposing enterprises to critique and blaming them for inappropriate use of their customers’ trust, but it is hard to find clear guidance on how companies should implement the Regulation. Multiple consultancy agencies offer their services stating that they can help organizations become compliant, however researchers in this field are still not certain of the best practices. Several privacy and data protection frameworks have been proposed in the last years (this will be discussed more in detail further in this chapter), roles

and responsibilities are being defined, stakeholders involved, however there is no "one size fits all approach" (*Chalcraft, 2018*). Despite all the recent advancements in analytics and big data technologies, there is a gap between what is possible for organizations and what is legally allowed; moreover, there is a lot of disagreement between the industry professionals regarding roles and responsibilities as well as methods of handling privacy concerns in information systems because after all, organizations are "just people", and the meaning and interpretation of the question "how?" will depend on the number of involved stakeholders and their goals (*Chessell, 2014*).

Another issue is that many organizations lack skills and expertise in implementation of regulations into business processes and tools, so they will face difficulties managing data under GDPR. The baseline for the discussion should be how to do the right thing for society, instead of how to avoid getting sued (*Greengard, 2018*). It appears that some might consider fines for non-compliance with the Regulation as "occasional losses" and treat them as a cost of doing business (*The Economist, 2019b*). Nevertheless, the issue of ethics in big data and analytics is being constantly raised and looked at from different perspectives. Tene and Polonetsky (*2013*) talk about serious concerns of big data, where amongst others, they point reader's attention to the problem of re-identification (discussed earlier in this chapter) of an individual by combining unrelated sources, issue of automated decision-making processes and predictive analysis and other difficulties of ethics of analytics. Authors discuss, for example, how predictive analysis can benefit society in cases such as planning disaster recovery in an earthquake prone area, but in other cases how personalized experiences can be rather intrusive. One of such examples is how by analyzing large datasets of purchase habits of consumers, the retail giant Target Inc. assigned a "pregnancy prediction score" to their customers, which was based on statistics of historical buying records of women who had signed up for baby registries, and then was able to predict a customer's pregnancy and even due date. Considering such examples and keeping in mind that different stakeholders will have different level of ethical boundaries, the question of an extent of privacy limitations remains open and requires further research.

Overall, the topic of privacy and ethics in information systems, and in particular tools and processes that are utilized for analytics, attracts attention of researches from multiple areas. Chen et al (*2012*) discuss the evolution of BI/BA, highlight key characteristics and capabilities of related tools and applications across various aspects of business and then review major emerging research topics across data analytics domain. Relational database management systems, data warehousing and its main concepts and data mining take top positions amongst foundational technologies proposed for research in data analytics, whereas privacy-preserving data mining is amongst the emerging research topics, thereby justifying the research question of this study.

2.5 Data Warehouse: Tool of Business Strategy

Modern organizations, whether small and medium companies or large multinational enterprises, manage large amounts of information on daily basis. Data that companies possess is one of the most important assets and in most cases, it has two purposes: operational and strategic (*Kimball and Ross, 2013, p.2*). Operational data allows flawless execution of day-to-day tasks, ensures quick and efficient execution of users' operations and resides in the operational systems such as order management systems, email servers and clients, customer service portals and support case management systems. In contrast, strategic data supports business decisions, provides overview of historical transactions, helps with forecasting, determining trends and predicting usage patterns; such data sets typically reside in DW/BI systems, which are also called Decision Support Systems (DSS). Inmon (*2005, p.15*) and Ponniah (*2010, p.13*) summarizes the differences between operational and DSS data as provided in the below table:

Authors	Operational Data	Strategic Data
I.	Used to run day-to-day operations	Calculated to meet the needs of the management
I.	Supports clerical function	Supports the managerial function
I.	Transaction-driven	Analysis-driven
I. & P.	Can be updated (access type: read, update, delete)	Can be recalculated, but not directly updated (access type: read)
I. & P.	Primarily current-value data, accurate at a time of access	Often – historical, archived, summarized data, values over time, snapshots
I. & P.	Repetitive procedures	Heuristic, ad-hoc requests
I. & P.	Large number of users, high access frequency	Relatively small number of users, low access frequency
P.	Optimized for transactions	Optimized for complex queries

Table 3: Operational and strategic data (Inmon, 2005; Ponniah, 2010)

It is not in the scope of this research to review detailed characteristics of operational systems other than databases. This study only differentiates between operational databases as OLTP (online transaction processing) systems and data warehouses as OLAP (online analytical processing) systems. In the course of this research, different concepts and characteristics of database systems are examined, so the key differences between OLTP and OLAP systems that are listed in the *Table 3* will be kept in mind.

Inmon (*2005, p.29*) offers the following definition of a Data Warehouse, which is widely used by other researchers in data analytics: "*A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions*".

DW characteristic: subject orientation

The "subject orientation" and application of data warehouse is defined by the type of the business and the industry. For example, for an insurance company, a subject may be car, life, health, travel; for retail industry – product, vendor, customer and sales. In the context of this research, the type of business investigated is online shop, where some of the major subject areas would be customer, sales, product, advertisement and promotion. Some of these areas require special attention due to application of PII protection. For example, customers would usually need to leave certain information about themselves to a company that is selling goods online as this information is required to successfully deliver customer's order and receive payment. These types of transactions require special attention and must be protected by appropriate technical means.

DW characteristic: integrated

Integration is related to the fact that DW pulls data from multiple sources – production operational databases, individual sets of data, archived data, external loads. Inmon (*2005, p.30*) argues that data integration is the most important aspect of DW as it must ensure data accuracy and consistency in naming conventions, measurement of attributes, encoding, and eliminate duplicates. If data is inaccurate and inconsistent, then by querying same information, different applications might return different results. *Figure 4* presents the issue of integration as described by Inmon (*2005, p.31*):

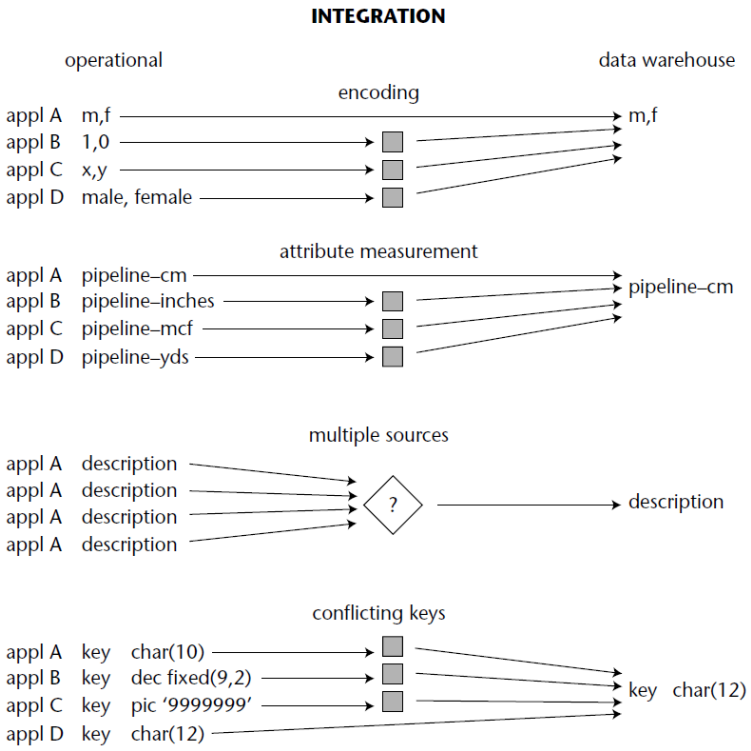


Figure 4: The issue of integration (Inmon, 2005, p.31)

DW characteristic: non-volatility

Unlike operational databases, where data is being updated regularly and incrementally, DW have their data loaded in a "snapshot" way and then accessed by running specific queries that will provide analysts with a report requested by management. Data is not usually updated, but rather new versions of data are added; when changes to data occur, they a written as a new snapshot, so that data can be kept historically, which typically helps in determining trends and building patterns. There are several GDPR-related concerns about non-volatility of DW; one such concern is "the right to be forgotten", since data is generally not deleted from DW, but rather moved to a different level of summary. Some of these concerns are addressed in the next chapter of this dissertation, combined with some applicable privacy design techniques and data protection frameworks.

DW characteristic: time-variance

As Ponniah (*2010, p.26*) describes time-variance: "*Every data structure in the data warehouse contains the time element*". Operational systems are designed to contain current or most recent data while DW contains historical data recorded in snapshots. Changes to data are tracked and recorded, so that analysts could identify trends in behaviours and make forecasts based on the patterns identified through historical records. Here comes another GDPR-related concern – "purpose limitation". By using historical data that was loaded to DW over a period of time, businesses might derive valuable information that other companies would not possess, which would eventually give them competitive advantage. However, if the reason for processing and use of data is different from the purpose for which it was collected, this is where privacy concerns are raised.

DW design approach: granularity

Although "granularity" is not included in the definition of DW, it represents an important concept of DW design. "Granularity refers to the level of detail or summarization of the units of data in the data warehouse" (*Inmon, 2005, p.41*). The more detailed the data, the lower level of granularity it has. Depending on the use case, different levels of granularity may be required, - for example, grouping by monthly sales amounts, weekly numbers of new customers or daily transactions. It may as well benefit analysis of business metrics measurements and key performance indicators. Granularity is usually decided based on the data types and the expected system performance for queries (*Ponniah, 2010, p. 28*). It may be argued that moving data to higher level of granularity could resolve several GDPR-related concerns, including purpose limitation and data minimization. It is not common to look at single records in DW, so privacy protection can be achieved by defining a purpose for processing summarized data and informing the users about how the new data is used.

DW design approach: partitioning

The purpose of data partitioning is to break it up into smaller and more manageable units compared to large datasets that take long time to process and analyze and consume a lot of physical resources (*Inmon, 2005, p.54*). In other terms, partitioning may also be called as grouping – each data unit belongs to a specific partition at each point in time. For example, one of the common groupings in data analysis would be by date, by business unit, by country or by type of product. While data partitioning is typically reviewed from the technical perspective and benefits such as performance improvement are outlined, it may be argued that partitioning could help with preserving privacy – let us say, if data about customers can be grouped by geographies, different processing activities would be allowed to be performed on sets of data containing information about different users, depending on the local regulations.

DW design approach: purging

Data purging is related to non-volatility attribute of DW and is one of the fundamental design issues. Inmon (*2005, p.64*) indicates that in some cases data is not purged from DW at all, but rather just "*simply rolled up to higher levels of summary*". There are several ways in which data can be purged or transformed, so this element should be an active part of the design process of DW (*Inmon, 2005, p.64*). Moreover, when talking about the "data erasure" principle of GDPR, then having a clear understanding of how purging process in a specific DW is defined, could help approach the compliance line.

While technical elements and characteristics of DW are important, one must not forget that the main purpose of this system is to serve the business. Before looking at the design and architecture of the system, everyone should start from defining DW project, involving relevant stakeholders and setting goals. Kimball and Ross (*2013*) not only present technical characteristics of DW, but also project them onto business requirements and describe roles and responsibilities of those involved in the DW planning and development process. It requires a lot of time, effort and resources, as well as detailed design and planning to build systems that would collect required information from multiple sources, and then store, process and extract data that would then assist the business with strategic decisions. All organizations face the task of data management at some point, and it is important to have the right tools and knowledge in order to do it right. While talking about the evolution of DSS, Ponniah (*2010, p.4*) notes that with the growth of businesses and competition in the 1990s, organizations became in desperate need of strategic information in order to achieve competitive advantage. Running just day-to-day operations was not enough, as companies needed different type of information. Therefore, data warehousing became more popular as a tool that was able to provide the data businesses were looking for.

2.6 Data Protection by Design in Information Systems

Privacy by design is not a new concept, though some might believe that it emerged with the development of GDPR. Early approaches to principles for guidance of data processing have their history back in 1970s (*ENISA, 2014, p.4*), and researchers have been investigating privacy concerns in information systems during the last few decades. In 1995, when Data Protection Directive was adopted by European Union (*Lord, 2018*), Ann Cavoukian developed and formalized her privacy by design approach to systems engineering, and within the next decades scholars had their attention on possibility to address data protection and privacy concerns in system development. In the course of this research, at least thirty academic publications as well as some online blog posts were reviewed on the topic of data protection and privacy by design in systems engineering, in particular with regard to protection of PII data and privacy concerns in BI/BA applications. Most of the reviewed resources were published after year 2000, so the privacy discussion in this literature is often based on the principles of both DPD and International Safe Harbour Privacy Principles[2] that were established in 2000. When in 2012 EC announced its plan to develop the GDPR, this raised a lot of grounds for discussion amongst researchers regarding the applicability and viability of the Regulation from the technical perspective. This study is focusing on the most recent literature published on the privacy topic within the last decade, and approximately 80% of the referenced sources about data protection and privacy were published after 2012. Some of the most commonly asked questions around this topic are about difference between DPbD and PbD, relation between privacy and security, conflict between individuals' and organizations' perspective, translation of legal obligations into business requirements and argument whether privacy can be "hardcoded". In terms of difference between DPbD and PbD, most researchers take them as one concept. Kroener and Wright (*2014*) distinguish between PbD, privacy by default and data protection by default and claim that PbD is implied by the GDPR, however not explicitly stated. Jasmontaite et al (*2018*) also separate DPbD and data protection by default and state that the concepts are interrelated, however DPbD typically refers to technical measures and safeguards of the application, while data protection by default refers to activation of policies and processes as organizational measures. In the context of this dissertation, PbD and DPbD are treated as one since DPbD implies PbD as used in the context of the Regulation.

In their study, Rommetveit et al (*2018*) find that one of the reasons why organizations are not promoting PbD is that data has too high value. Application of privacy design techniques would force them to collect less data, and this would have negative impact on business.

[2] Developed between 1998 and 2000 in order to prevent private organizations within the European Union or United States which store customer data from accidentally disclosing or losing personal information (https://en.wikipedia.org/wiki/International_Safe_Harbor_Privacy_Principles)

One of the biggest concerns regarding implementation of privacy techniques is that it brings challenges and forces organizations to reevaluate their data management activities. It is stressed by multiple researchers that companies are not trained or geared towards considering users' privacy concerns and they lack understanding of such data operations as collection, storage and processing that are taking place within organization (*Rommetveit et al, 2018*). Koops and Leenes (*2014*) argue that businesses "*have little clue how they should go about 'designing in' privacy*". In the context of data warehousing, this creates problems because if organizations do not always know how much data they have, where it is stored and who is responsible for it, then it means they are not completely in control of the resources they possess, which leaves them vulnerable to potential data breaches.

One of the frameworks most commonly referenced in the literature on privacy topics is Cavoukian's (*2009*) foundational principles of PbD. Below table summarizes the framework:

PbD Principle	Description and explanation
Proactive not Reactive; Preventative not Remedial	Prevent privacy invasive events, commit to highest standards of privacy, recognize poor privacy designs
Privacy as Default	Purpose specification, collection, use, retention and disclosure limitation, data minimization
Privacy Embedded in Design	Privacy an essential component of the core functionality
Full Functionality – Positive-Sum, not Zero-Sum	Accommodate all legitimate interests and objectives, do not put privacy against security – implement both
End-to-End Security – Lifecycle Protection	Secure lifecycle management of information, assure confidentiality, integrity and availability of data
Visibility and Transparency	Accountability, openness, compliance
Respect for User Privacy	Consent, accuracy, access, compliance

Table 4: Foundational principles of PbD (Cavoukian, 2009)

It is suggested that this framework should apply to every stage of systems development process: "*Privacy must become integral to organizational priorities, project objectives, design processes, and planning operations*". However, as discussed earlier in this chapter, many authors criticize it for being too vague, unrealistic, focused only on individuals and not applicable to some industries. Koops and Leenes (*2014*) state that it is still not clear how legal obligation to ensure PbD should be implemented in practice. Cavoukian's framework might seem to be too user-centric and do not take requirements of business into consideration, however with DPbD being a legal requirement as per GDPR, organizations cannot afford questioning requirements to implement privacy in their systems and processes. If business is not afraid of fines for being non-compliant, they should be afraid of losing trust of their customers and partners (*Rommetveit et al, 2018*). One of the

challenges of DPbD is combining legal principles and engineering since legal texts can have different interpretations, and technical requirements are typically non-ambiguous; however, Rommetveit et al (*2018*) suggest that this may be resolved by "*applying generic legal principles to concrete technological contexts*" in design processes, and also "*enhance oversight of decisions*" in procedural checks.

While there might be no complete and comprehensive framework that would cover all aspects of privacy (*Blix et al, 2017*), many researchers take an opportunity to approach this problem creatively. Nowadays, the issue is investigated from different perspectives, and researchers are creating frameworks that can help organizations implement privacy techniques. Spiekermann and Cranor (*2009*) separate the concepts of "privacy-by-policy" and "privacy-by-architecture" and review them from both technical perspective and economic feasibility. Authors also suggest that companies should keep in mind users and their perspective of what privacy breach means. All information systems typically perform some data-related tasks (transfer, storage, processing), therefore users become concerned about privacy when collected data is no longer under their control. This concern also proves the relevance of this study – DW systems are designed to store historical data from multiple sources, therefore these systems are far beyond user's control.

Spiekermann and Cranor (*2009*) also believe that developers hold a major responsibility for privacy engineering, as they are the ones that actually design, architect and develop the systems. However, it is interesting to find that developers themselves might have different perception and interpretation of PbD practices. Hadar et al (*2017*) have interviewed 27 developers from different domains who practice software design and investigated their point of view regarding implementation of PbD to assess its viability. It was found that not only were the respondents unsure of the meaning of privacy, but also questioned feasibility of its implementation. It is noted that in the "real world" business considerations are treated with higher priority over users' privacy, and participants were discussing privacy as a social concern based on norms of morality and ethics rather than a technological concern.

While some researchers argue that privacy is too vague and cannot be aligned with specific engineering requirements (*Van Dijk et al, 2018*), others propose their solutions to this problem. Hoepman (*2019*) offered a framework how PbD philosophy can be applied in system design. Back in 2012, he offered eight privacy design strategies that can cover all concerned FIPs. The strategies are grouped into two classes: data-oriented and process-oriented. Data-oriented strategies correspond to the concept of "privacy-by-architecture" described by Spiekermann and Cranor (*2009*), and the process-oriented strategies refer to the "privacy-by-policy" approach. Data-oriented strategies are: minimize, hide, separate and aggregate; and process-oriented ones are: inform, control, enforce and demonstrate. To

test application of these principles, Hoepman et al (*2014*) defined legal requirements that these design strategies cover and noted that not every legal requirement can be implemented by technical means, though in that case organizational measures can apply (policies and processes). He then mapped derived privacy design strategies to legal requirements and listed which principles cover what requirements and to which extent.

	Purpose limitation	Data minimisation	Data quality	Transparency	Data subject rights	The right to be forgotten	Adequate protection	Data portability	Data breach notification	(Provable) Compliance
MINIMISE	o	+								
HIDE		+					o			
SEPARATE	o						o			
AGGREGATE	o	+								
INFORM				+	+				+	
CONTROL			o		+			+		
ENFORCE	+		+			+	+			o
DEMONSTRATE										+

Legend: +: covers principle to a large extent. o: covers principle to some extent.

Figure 5: Mapping PbD strategies with legal requirements (Hoepman et al, 2014)

The above mapping in *Figure 5* is referred as "Hoepman's taxonomy of privacy design strategies" or simply "Hoepman's taxonomy" further in this dissertation. The taxonomy is especially relevant to this study since for testing the framework Hoepman et al (*2014*) take an abstract storage model as a point of departure and apply eight privacy design strategies, which can be translated to the process of data management within DW as well (*Figure 6*).

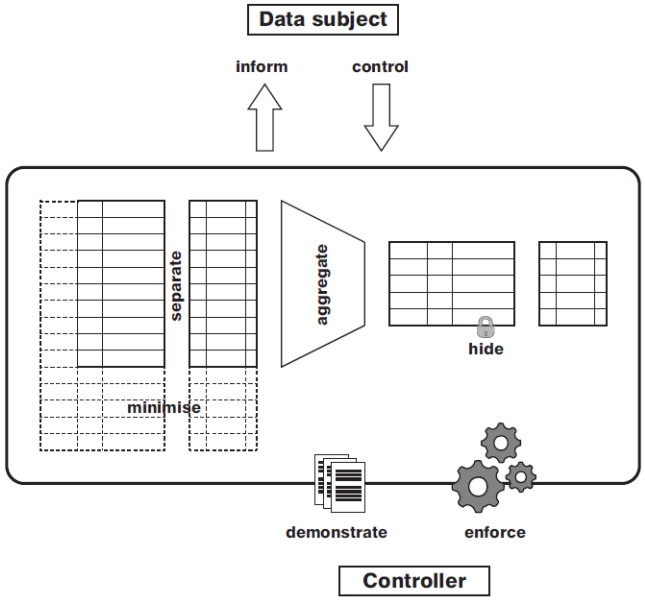


Figure 6: Privacy design strategies applied in a database system (Hoepman et al, 2014)

Koops and Leenes (*2014*) argue that privacy cannot be implemented by technical means, but rather must be promoted with the right communication, processes and mindset ("soft-coded"). They conclude that "*there are simply too many complications for data controllers to be able to effectively implement 'hard privacy by design'*", so PbD cannot become a requirement for systems development and the focus should be on creating privacy mindset rather than hardcoding privacy. Accordingly, Blix et al (*2017*) propose a framework for translation of data protection into business requirements using design science methodology. Unlike Hoepman's framework, Blix et al (*2017*) address all phases of system development process, starting from preparation and collection of business requirements, assessment of factors to consider and actual implementation. As a result, one should receive GDPR-compliant system designed according to PbD principles. This framework will be used as a foundation for the DW design process in this study (*Figure 7*).

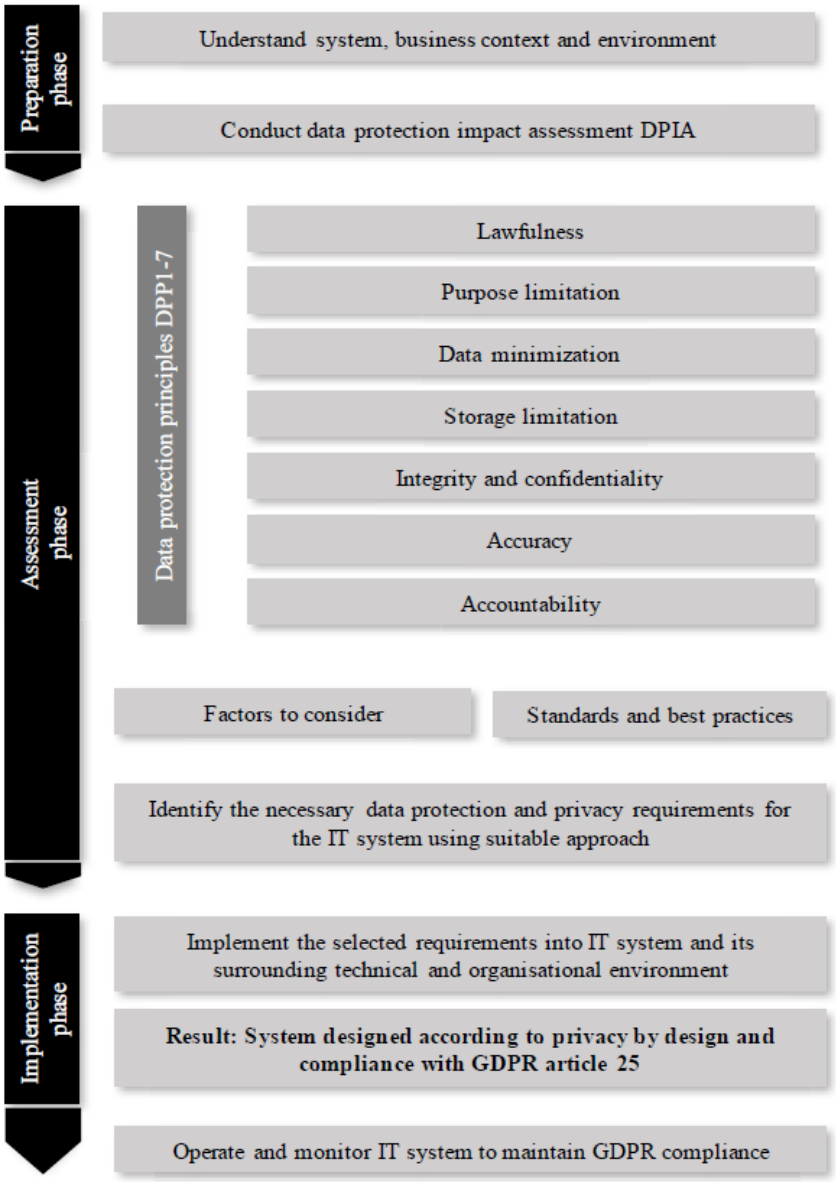


Figure 7: Framework for PbD (Blix et al, 2017)

2.7 Summary

This chapter provided extensive review of literature on the topics of privacy, data protection, GDPR, data analytics and data warehousing, discussing main concepts presenting major arguments. The background and implications of the GDPR for both individuals and organizations as well as definitions and core principles were outlined, and importance of the Regulation examined. Also, the concept of PII data was reviewed and explored in the context of the GDPR. It was discovered why PII data requires special protection, how organizations collect and use it, and which threats companies and individuals should be aware of. In terms of lifecycle of data in organizations, it was discussed how data is valued as an asset and how it can help companies make strategic decisions. The arguments regarding data protection were described with tools and techniques used in data analytics, one of which were data warehouses. Core DW characteristics and design techniques were reviewed, for which some of the GDPR-related concerns were raised and discussed. Proposed solutions for these concerns are addressed later in the next chapters. And finally, this chapter was finalized by examining some relevant DPbD and PbD principles, standards, design patterns, techniques and frameworks, which will be later investigated on the subject of applicability in design and architecture of DW systems.

3 Research Methodology and Design

3.1 Research Approaches in Information Systems

Information Systems (IS) research differs from other, more traditional types of researches, as it does not just investigate technology or social science alone. Discussions framed within IS discipline are appealing to students and academics because of their interdisciplinary nature. It is rare that IS research would only focus on one aspect of a particular area – for example, computer science studies often focus on the development or architectural side of systems, where the final product would be a prototype of a system or a module. Students who pursue their degree in arts are often expected to generate an idea or create a piece of art. Social studies are known to utilize surveys and interviews in their research, and then make conclusions based on the analysis of findings. However, IS discipline is contributing to the academic community with its research focused on emerging technologies and by addressing problems from other domains, combining research techniques. Truex et al (*2006*) investigate the question whether IS field can only borrow theories from reference disciplines, or whether it can contribute to them back with innovative solutions. IS research may distinguish between several types of theories: for analyzing, for explaining, for predicting, for explaining and predicting, and for design and action. All these types of theories are used as means of growing knowledge in a given field (*Trux et al, 2016*). Decision Support Systems were known to be amongst top areas of interest for research in the 1990s, where architecture development, effective use of data and improvement of IS strategic planning were some of the key IS management issues as perceived by IS executives in the US (*Galliers, 1995*). With the changes in technology, the above topics need to be reviewed from new perspectives – considering that amount of data collected by organizations globally is only increasing, this creates new scopes for research.

This chapter provides an overview of approaches in IS research and then outlines the relevant methods and techniques used in the course of this study. It also justifies the frameworks and design techniques that were chosen to answer the research question and prepares the foundation for the next chapter, where application of the design study is tested and described on an example of a DW system for an e-commerce website.

There are multiple approaches that are applicable to IS research, however depending on the topic and area, some methods and techniques are more appropriate than others. Galliers (*1995*) proposes a framework to aid choice in IS research, and according to his framework, some of the most relevant approaches are field experiment, survey, simulation and review. Some less common, but still applicable approaches can also be case study, futures research and lab experiment. It can be argued that IS research community is moving away from focusing on more technical issues and tends to explore behavioural problems

(*Parker et al, 1994*), which is reflected in this study as well. GDPR is often discussed in researches from the perspective of regulation and policy, while in its nature it affects IS and technical measures for their implementation.

It was decided that this research would be qualitative, and *design science* was chosen as a methodology for this study as the one that most appropriately reflects the nature of performed activities. Design science is a problem-solving paradigm that is both a process and a product, and it seeks to define ideas, practices and technical capabilities that would resolve organizational issues effectively in a creative way (*Hevner et al, 2004*). *Figure 8* represents relations between business, organizational, IT and IS elements of strategy and architecture, and effective transition between them requires extensive design activities.

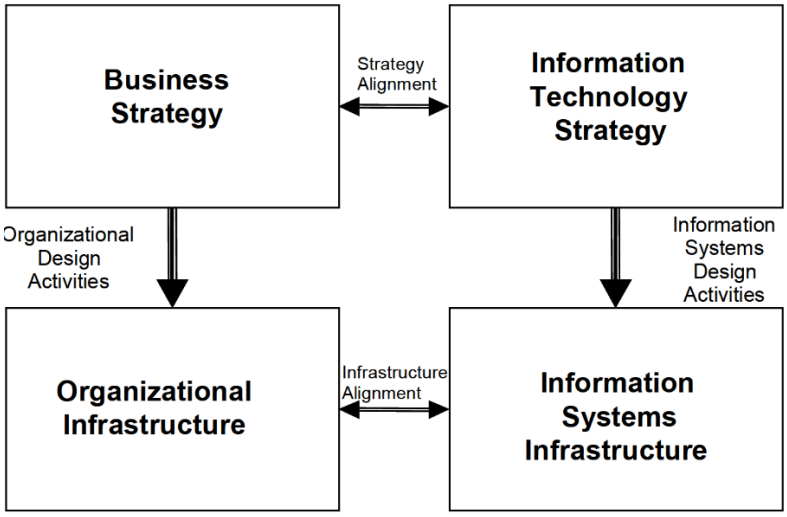


Figure 8: Organizational and IS design activities (Hevner et al, 2004)

Vaishnavi et al (*2004/17*) state that design science complements positivist, interpretive and critical perspectives and involve two primary activities: "*creation of new knowledge through design of novel or innovative artifacts (things or processes)*" and "*analysis of the artifact's use and/or performance with reflection and abstraction*". The methodology fits this study perfectly as the goal of research is to create and outline a process of designing and architecting a DW system with implemented privacy techniques. The structure of the process should combine best practices and known frameworks from disciplines that are not directly related. Offermann et all (*2009*) outline the following steps of the design science research: problem identification (includes literature review), solution design, evaluation and results summary. In this study, the problem is identified in *Introduction* and *Literature Review* chapters, solution design proposed in *Research Methodology and Design*, and then results are described and evaluated in *Framework Analysis and Application* chapter with conclusions summarized in the last section of this dissertation.

3.2 Choosing Privacy Framework for Design

Earlier in this chapter, several frameworks for implementation of privacy were discussed. For the solution proposed in this research, frameworks presented in *Figure 7* and *Figure 8* are utilized: Hoepman's taxonomy from the design and architectural standpoint and PbD framework of Blix et al (*2017*) from the organizational perspective.

Art. 25 of the GDPR is one of the core articles of the Regulation as it lists the main principles of PII data processing: *lawfulness*, *fairness* and *transparency*, *purpose limitation*, *data minimization*, *accuracy*, *storage limitation*, *integrity* and *confidentiality*, and last one that ensures compliance with all the other principles – *accountability*. Blix et al (*2017*) place them in the context of PbD and call them Data Protection Principles (DPP). Hoepman et al (*2014*) address all of these principles (and also add two more: *data subject rights* and *the right to be forgotten*) by mapping them to privacy design strategies that are believed to resolve privacy concerns in systems design: *minimize*, *hide*, *separate*, *aggregate*, *inform*, *control*, *enforce* and *demonstrate*. In the previous chapter, other sources that refer to privacy principles of GDPR were also reviewed, and this was outlined in *Table 2*. To ensure that all areas are covered, each principle will be marked with DPP# notation proposed by Blix et al (*2017*), as this will help optimize the finalized privacy framework proposed in this study. *Table 5* labels each category of privacy principles as they appear in the referenced sources.

DPP#	**GDPR**	**Hoepman et al (*2014*)**	***Table 2* from the *Literature Review* chapter**
DPP1	Lawfulness, fairness and transparency	Transparency	Openness, transparency and notice; Consent and choice
DPP2	Purpose limitation		Purpose legitimacy and specification
DPP3	Data minimization		
DPP4	Storage limitation	Data portability	Collection and retention limitation
DPP5	Integrity	Adequate protection	Information security and protection
DDP6	Confidentiality	Data breach notification	Use and disclosure limitation
DPP7	Accuracy	Data quality	Accuracy and quality
DPP8	Accountability	Compliance	Accountability and privacy compliance
DPP9	Art.15 Right of access […]	Data subject rights	Individual participation and access;
DPP10	Art.17 […] erasure	The right to be forgotten	

Table 5: Labeling privacy principles with DPP notation (proposed by author of this dissertation after aggregating multiple frameworks)

While each of the principles labeled in *Table 5* is addressed in Hoepman's taxonomy, the framework is not yet complete. It is not enough to just list design strategies and claim that their implementation will ensure that the system would be GDPR-compliant. It is also important to categorize these principles as ones that refer to either organizational or technical measures to be applied by organizations. As suggested in previous chapter, privacy design strategies can be categorized as data-oriented and process-oriented, which corresponds to "privacy-by-architecture" and "privacy-by-policy" approaches respectively as described by Spiekermann and Cranor (*2009*). Following analysis of methods proposed for PbD application in IS design, the summarized framework was created in *Table 6* that reflects which privacy design strategy applies to which data protection principle and whether they are data-oriented or process-oriented. Using recommendations of Blix et al (*2017*) suggestions are derived for each DPP regarding applicable implementation measures.

	DPPs	**PbD Strategies**	**Organizational Measures**	**Technical Measures**
Data-oriented / Privacy-by-architecture	*DPP2*	Minimize, separate	Strategy, policies, processes	Data inventory, meta-data, tagging, reporting
	DPP3	Minimize, hide, aggregate	Strategy, policies, processes, IAM	Centralized storage, proxies, pseudonymization, stripping
	DPP5	Hide, separate, enforce	IAM, encryption, physical security	End-to-end encryption, data validation
Process-oriented / Privacy-by-policy	*DPP1*	Inform	Strategy, policies, processes, legal measures	Embedded transparency, embedded legal measures, non-repudiation services
	DPP4	Control	Awareness, data lifespan	Tractability, self-wiping, reporting
	DPP6	Inform	IAM, encryption, key management	End-to-end encryption, authentication, authorization
	DPP7	Control, enforce	Data completeness awareness, data management, data normalization	Input validation, data dispute handling, data cleansing
	DPP8	Enforce, demonstrate	Strategy, policies, standards, awareness, certifications	Authentication, authorization, audit trails, monitoring, data loss prevention
	DPP9	Inform, control	Policies, data management	Tractability, reporting
	DPP10	Enforce	Policies, processes	Self-wiping, automation

Table 6: Categorization of PbD strategies and DPP (proposed by author of this dissertation after aggregating multiple frameworks)

3.3 Choosing DW Architecture

Research articles on the data warehousing topics usually refer to William H. Inmon and Ralph Kimball as authors whose architectures of DW systems are most commonly used across enterprises. While both authors describe dimensional nature of DW (the concept of dimensions is described later in this chapter), the main difference between the two offered architectures is the approach to building the system: *top-down* and *bottom-up* (*Poniah, 2010, p.20; Luján-Mora and Trujillo, 2004*). Inmon is one of the advocates of *top-down* approach that states that DW is a product of an enterprise, where data is stored for the whole organization, and then separate data marts are derived from the central repository to fulfil the needs of individual teams, groups or departments. This may benefit organization in a way that if several business units require to query similar information, there would be less chance of having discrepancy in finalized results since the data is pulled from one true source. Kimball, in contrary, is a leading proponent of the *bottom-up* approach in DW architecture, which is the opposite to the one of Inmon's. This approach offers to build individual data marts to fulfil analytics and reporting needs of separate business units within enterprise first, and then expand by adding, integrating and merging these data marts in small iterations to present them as a consolidated enterprise data warehouse.

While both approaches have their benefits and drawbacks, another aspect of given architectures was observed in the course of this research – compatibility with DPbD principles and potential applicability of PbD philosophy to each stage of DW lifecycle. It was discovered that Kimball's approach appears to be more friendly to application of privacy and data protection techniques. His main thesis regarding the purpose of DW is that "*first and foremost, the DW/BI system must consider the needs of the business*" (*Kimball and Ross, 2013, p.1*), and in the book he actively promotes the idea of "business requirements first", which states that organizations should know exactly the questions their DW needs to answer before the planning phase of a DW project begins. This approach resonates with requirements set in art.25 of the GDPR as it allows to define, limit and minimize the amount of data that needs to be collected at the earliest planning stage of system development process. In contrary, Inmon (*2005*) states that development of DW is the opposite to the classic traditional "waterfall" SDLC (where requirements are gathered, reviewed and understood first, and each next task of the process is triggered only when previous task is finished). He suggests that DW is built around the data that is already available and which is then integrated and tested, so systems are built on data. This approach is later revisited throughout the book and backed up by multiple statements. For example:

"*Organizations may build a data warehouse for one purpose, and then discover that it can be used for many other kinds of DSS processing*" (*Inmon, 2015, p.42*)

"*The granular data found in the data warehouse is the key to reusability, because it can be used by many people in different ways*" (*ibid, p.42*)

"*Perhaps the largest benefit of a data warehouse foundation is that future unknown requirements can be accommodated*" (*ibid, p.43*)

This approach at its nature goes against principles of data protection by design and by default, especially if we consider reusing collected data for different purpose from the one it was originally collected for. It was concluded that Kimball's approach to DW design and development is more suitable for the purpose of this research, so the subject of applicability of DPbD and PbD frameworks is analyzed and tested against best practices of Kimball's DW/BI architecture. Moreover, in chapter 21 "Big Data Analytics" of their book Kimball and Ross (*2013, p.541*) provide suggestions regarding governance best practices for big data and stress the importance of privacy from governance perspective, stating that one should not choose big data over governance. Nevertheless, work of Inmon is also referenced throughout this study in cases where description of certain core DW concepts is found more appropriate and easier to understand.

Relational Database Modelling

As it was mentioned in the previous chapter, DW systems as well as many operational databases have dimensional nature (*Kimball and Ross, 2013, p.7*). One can think of a DW as a collection of tables, each consisting of rows and columns (relations). Each row would typically represent a single record in the database – a customer, an order, an employee; and columns would represent relevant attributes of each record. For example, some of the attributes of the "customer" can be name, last name, email address, billing address, etc. The difference between a DW system and an operational system in this sense is that each row in operational system would only represent the current state and have the lowest level of granularity. For example, "products" table in operational database will contain information about every item available for sale in the store at the moment, but will not have details about items that were available on the same date last month; in contrast, "products" table in the DW will contain data about products that were available for sale at any point in time, providing additional dimensions of this data with monthly or quarterly granularity (i.e., showing growth in products number over time). Dimensional models can be referred to as either *star schemas* or *online analytical processing (OLAP) cubes*, depending on whether a database is relational or multidimensional respectively (*ibid, p.8*). Logical design of star schemas and OLAP cubes is similar, and they mostly differ in physical implementation; regarding differences from PbD perspective – OLAP cubes offer better security in terms limiting access to detailed data and providing more options to work with summarized data. The solution proposed in this research suggests using both models, where most detailed

information is loaded into a start schema and later populated into OLAP cubes using aggregation methods to summarize and group the data, serving as additional layer of data protection and access limitation. For reference, Figure illustrates star schema and OLAP cubes as presented by Kimball and Ross (*2013, p.9*).

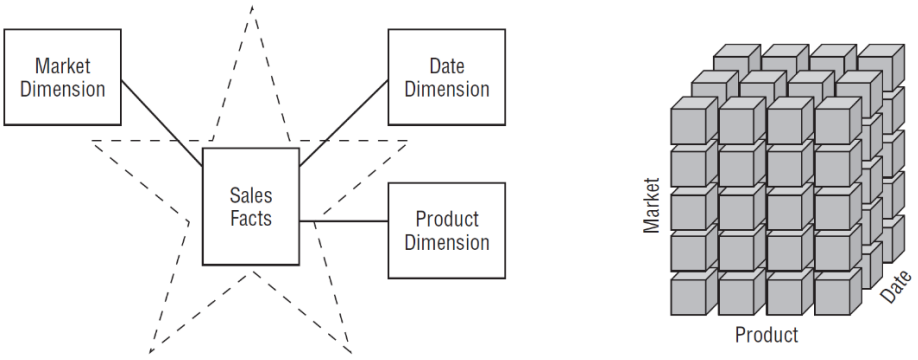


Figure 9: Star schema vs OLAP cube (Kimball and Ross, 2013, p.9)

Facts and Dimensions.

Other elements of relational databases that one must be familiar with are *facts* and *dimensions*. This study will not focus on technical details, so these terms are only explained on a high level to ensure this contributes to understanding of the proposed solution. As discussed earlier, one can think of a DW system as a collection of tables with rows and columns. Each table is typically characterized as a *fact* or a *dimension* table. A fact table stores low-level measurements resulting from a single business process, where each row represents a measurement event, and the data of each row has specific *grain* (*ibid, p.10*) such as one row per item sold during specific transaction or one row per customer registered with unique email address on the online website. Example of *fact* can be euro sales amount. *Dimension tables* complement fact tables and contain the context associated with the measurement event; they often have many attributes (columns), fewer rows than fact tables and are defined by a single primary key (PK), which allows to join related fact tables.

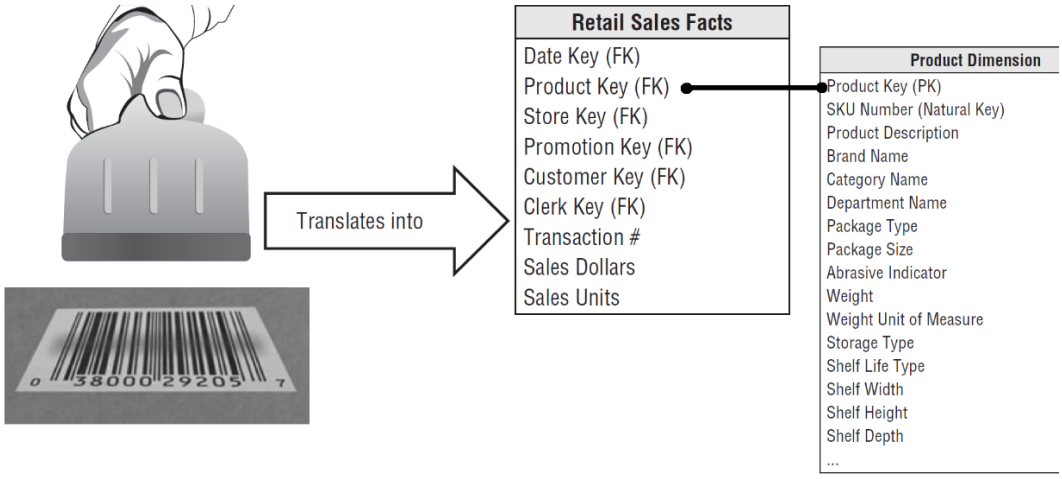


Figure 10: Dimension and fact tables relation (Kimball and Ross, 2013, pp.11,13)

Kimball's DW/BA Architecture

Four core elements of Kimball's DW architecture are *operational source systems*, *ETL system*, *data presentation area* and *BI applications* (*ibid, p.18*). For the purpose of this research it is important to understand the difference between them and strategic significance of each. *Operational source systems* capture business transactions and store operational data (see *Table 3*). In the context of DW, there is little or no control over the content and the format of the data that is retrieved from operational source systems, and their purpose is to make data available for further extraction and load into DW. These systems help performing day-to-day tasks and only store current data. *Extract, transformation, and load (ETL) system* is every work area, data structure and set of processes between the operational sources system and the DW presentation area. ETL system is responsible for extraction and reading of source data, and this is when data starts belonging to the DW (*ibid, p.19*). In the context of this study, ETL system is the core element of DW design since this is where all the data manipulation and transformation take place before the process rolls over to the presentation area. *Data presentation area* is where data is organized, stored, and made available for access and querying by analysts or BI applications (*ibid, p.21*), and this is what business can see. Star schemas and OLAP cubes belong to the presentation area of DW, which should be structured around business requirements – one should have a clear idea of questions that DW should answer (this resonates with the purpose limitation privacy strategy). And finally, *BI applications* refer to tools used to query data from DW, pull reports and visualize them in appropriate format. Improved decision making is the whole purpose of querying data from DW (*ibid, p.23*).

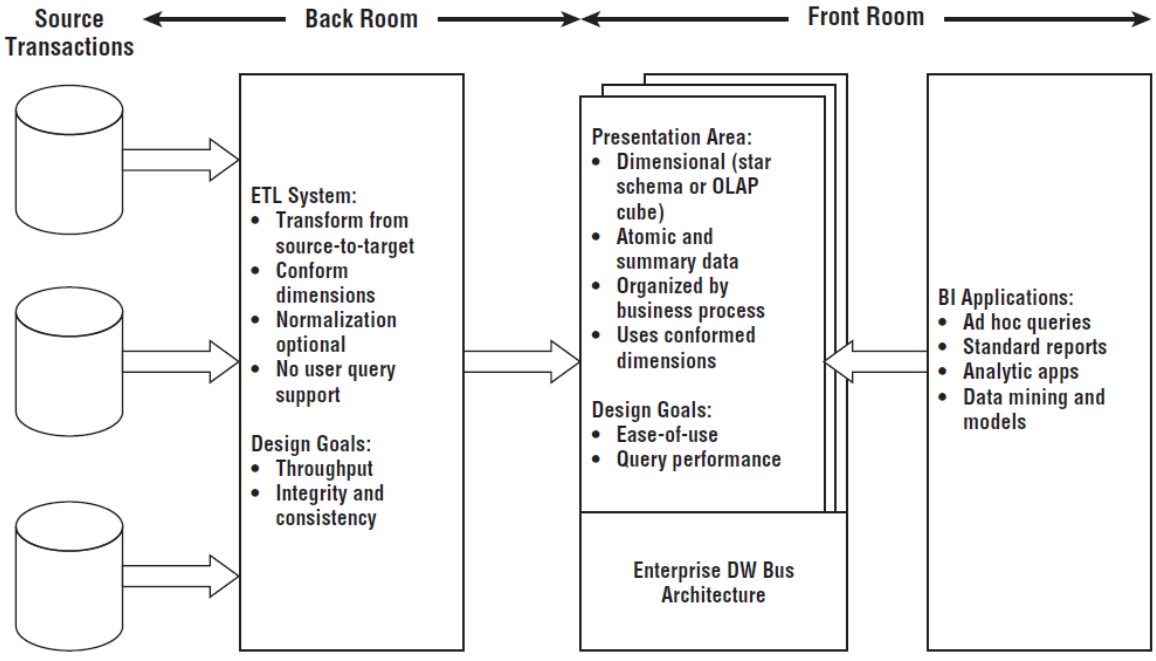


Figure 11: Elements of Kimball's DW/BI architecture (Kimball and Ross, 2013, p.19)

3.4 DW Design Methodology and Requirements

As in this research the privacy framework is tested on the design process of DW for an e-commerce website (online shop), it is first worth explaining the overall architecture of the website in order to show where exactly DW fits in. An average e-commerce website is built using multi-tier architecture that consists of the following layers: *presentation*, *application* and *data* layer (*Nagaty, 2010*). *Figure 12* provides a high-level interpretation of the three-tier architecture. Depending on the implementation, some layers will be either merged or separated. *Presentation layer* typically refers to the client, which can be a browser or an application on a mobile device or tablet. *Application layer* represents the server side: web-server that is handling front-end communication with the client and application server (back-end calculations and processing). In some interpretations, presentation layer refers to the web-server, and application layer – to the application server accordingly. However, in most implementations, data layer is separated as it typically corresponds to operational database systems that are handling everyday transactions of the application (e.g., new user registrations, online purchases, subscriptions, account updates, etc.)

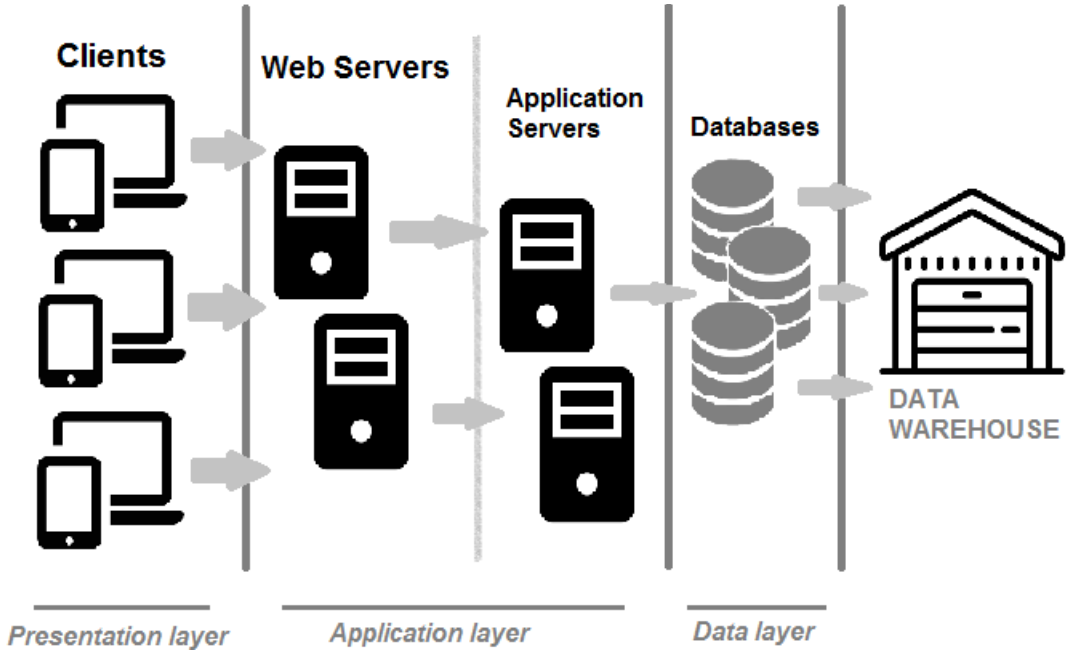


Figure 12: Three-tier architecture of e-commerce website

As discussed in the previous chapter, DW system is developed when business needs to find answers for strategic questions. DW is never customer-facing and is typically intended for internal use only. If organization decides to outsource business analytics, in that case the outsourced company will be considered a data processor in the context of the GDPR, while the business that provides the source data would only be considered a data controller. Regardless of the implementation or the business model, the GDPR principles are applicable in both cases.

Kimball's DW Lifecycle approach is presented in *Figure 13*. The flowchart illustrates tasks, their order, dependencies and concurrencies. Depending on the scope of the project, each phase might take different amount of time and resources. Kimball's approach addresses planning, design, development, deployment and growth of DW and shares principles of agile methodologies: focus on business value, collaboration with the business, and incremental development. Unlike Inmon's architecture, Kimball and Ross (*2013*) are better in projecting technical characteristics and capabilities of DW onto business requirements and processes, describing everyone's roles and responsibilities not only in terms of their every day job, but how they impact data management strategy overall.

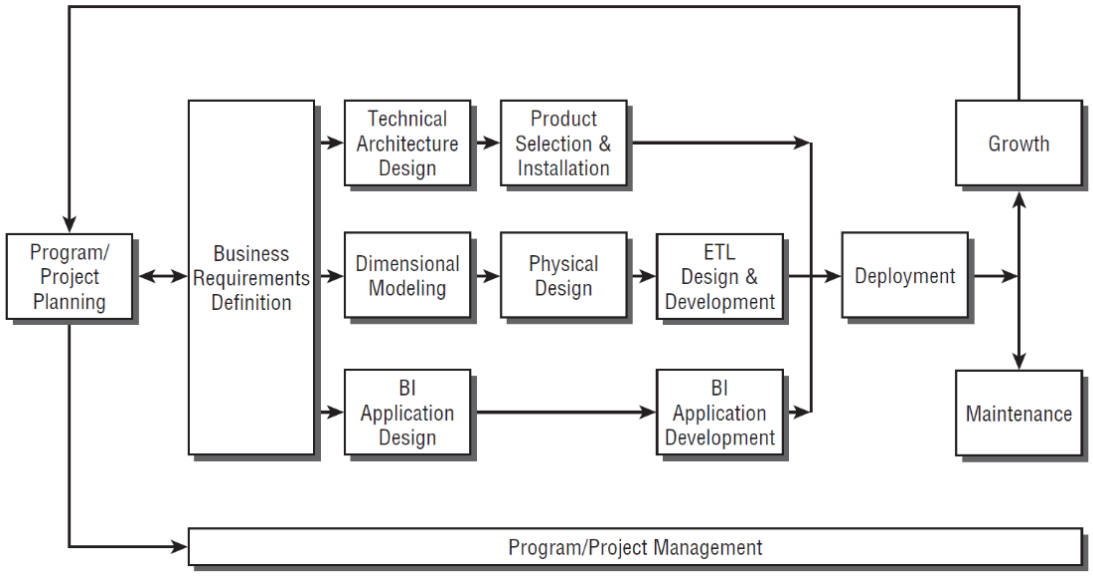


Figure 13: Kimball's DW Lifecycle Diagram (Kimball and Ross, 2013, p.404)

It is important to keep the GDPR principles in mind at every step of the outlined process, so one of the tasks for this study is to identify which DPPs should be implemented at which stage of the DW lifecycle and how – this is discussed in the next chapter.

Kimball and Ross (*2013, p.37*) suggest that following fundamental concepts and techniques must be followed during dimensional design process:

- Gather Business Requirements and Data Realities
- Collaborative Dimensional Modelling Workshops
- Four-Step Dimensional Design Process
    - Select the business process
    - Declare the grain
    - Identify the dimensions
    - Identify the facts
- Star Schemas and OLAP Cubes

Song and LeVan-Shultz (*1999*) outline this process with the flow illustrated in *Figure 14*.

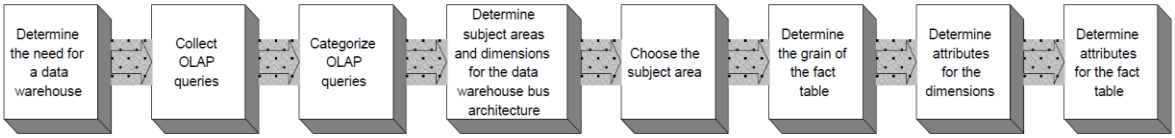


Figure 14: Dimensional modeling for e-commerce DW (Song and LeVan-Shultz, 1999)

In terms of gathering business requirements, it was mentioned earlier that organizations must have a clear understanding of the need for DW and which business questions the system should answer. On the operational side, all information and reports that users are getting out of the DW are derived by querying the tables available in the system using **SELECT** statements – whether by writing ad-hoc queries each time report is needed or by interacting with the DW via BI application. In the case of BI applications, the queries are pre-coded in its functionality. Each query that is run against the DW can be translated into a simple human-readable question that will represent which information we are looking for in the DW. Though in the real world it is usually vice versa – questions are translated into OLAP queries that are run against the DW tables pulling the required information. For example, let us say management of sales department needs to know last month's revenue gained specifically from selling their new product called "ABC" broken down by each EU country, and let us assume that information is pulled from the table "`Orders`" that contains all needed columns, and no joins are required with other tables. *Table 7* represents an example how the table could look like and which information it would contain. There typically would be more column in the table, but for simplicity only the relevant ones are provided.

order_id	order_date	product_name	charge_amount	country
123456	2019-03-05	ABC	€99.90	DE
234567	2019-03-06	DEF	€98.60	IE
345678	2019-03-07	GHI	€99.80	FR
456789	2019-03-08	JKL	€96.50	IT
…	…	…	…	…

Table 7: Sample "Orders" table

So, the OLAP query to pull required information from *Table 7*, considering the question that needs to be answered, could look like:

```
SELECT product_name, SUM(charge_amount) AS "Total", country

FROM Orders

WHERE product_name = 'ABC' AND order_date = '2019-03'

GROUP BY country
```

The output of this query would depend on the number of lines and look like *Table 8*.

product_name	total	country
ABC	€19,980.00	DE
ABC	€14,790.00	IE
ABC	€11,976.00	FR
ABC	€28,950.00	IT

Table 8: Output from "Orders" table received as a result of running a query

Song and LeVan-Shultz (*1999*) in their study of DW design for e-commerce website provided a comprehensive list of almost 100 questions that management of the business may be asking and categorized them depending on the line of business (department). Creating a list of such questions is one of the first steps of the DW project implementation as this helps to clarify the purpose and minimize the risk that unnecessary data will be collected and processed. Song and LeVan-Shultz's (*1999*) list is complete enough to justify the need for the DW, however considering that the study was conducted twenty years ago, no privacy concerns were addressed, and nowadays many of the questions would not pass the "purpose limitation" criteria. While it is true that the more questions are written down, the better idea it gives about the needs of the business, it is important to remember that the task for the project team is to keep privacy considerations in mind, so another step in the process would be to review the list, justify the questions asked and remove or re-evaluate and replace the questions that might interfere with privacy. It may seem like this exercise does not benefit the business, restricting companies and potentially putting them in disadvantage in the competitive race, however compliance with regulations and saving customers' trust eventually retains costs and reputation. All questions for OLAP queries that were proposed by Song and LeVan-Shultz (*1999*) were analyzed and tested against the privacy principles. *Appendix 4* outlines these questions and marks them as either *fully ("Yes")*, *partially ("Caution")* or *not ("No")* compliant.

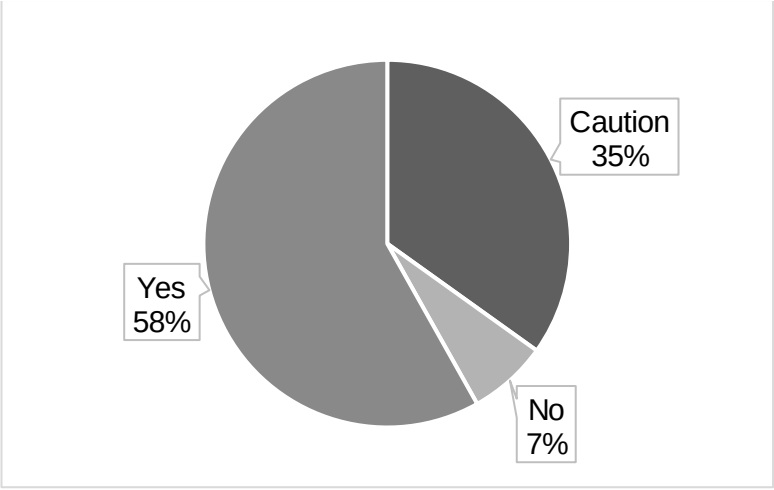


Figure 15: Compliance % of OLAP queries against DPPs

One must be cautious with partially compliant questions. Though they can be resolved by applying appropriate privacy strategies, they can potentially interfere with privacy considerations. Non-compliant questions directly involve PII data processing, so the business must either decline or replace them. As illustrated in *Figure 15*, approximately 42% of questions that businesses might need answers for, interfere with privacy concerns. It may be argued that the analysis of Song and LeVan-Shultz's (*1999*) questions conducted in this study was a subjective process. Therefore, it is suggested that this exercise should be conducted during the DW project planning phase and involve all required stakeholders.

3.5 Project Planning, Defining Roles and Responsibilities

As any other project in an enterprise, DW development process must follow a specific flow. First and foremost, before the planning starts, the business must understand their need for a DW. Kimball and Ross (*2013, p.404*) suggest that the following stages must be completed before gathering the requirements and architecting technical and organizational design of the system: *assessing readiness*, *scoping and justification*, *staffing,* and *developing and maintaining the plan*. The most critical tasks are to get support of strong executive business sponsor, have a valid business motivation for starting DW/BI project and evaluate feasibility of resources and data (*ibid, p.407*). The cost and time of implementation must be also considered, so if required information is already available in operational systems, then it may be argued that there is not enough justification for building a DW system. Another important step is to define the scope of the project. Many DW projects fail because business simply decided to implement everything at once (*Inmon, 2005, p.269*). To build a successful DW, there must be support from both business and technical teams. From the business side, besides the sponsor, *business driver*, *lead* and *users* must be involved as well (*Kimball and Ross, 2013, p.408*). The DW cannot be developed successfully without involving these parties since the whole purpose of the system is to support the business itself. One must as well involve "hybrid" representatives who either hold technical roles and understand the business, or work with the business teams but understand the technology; amongst others, these could be: *business analyst*, *data steward* and *BI application designer/developer* (*Kimball and Ross, 2013, p.408*). Amongst the resources that must be dedicated to the project from the IT/IS department, most commonly involved roles are *project manager*, *technical architect*, *data architect/modeler*, *database administrator*, *metadata coordinator*, *ETL architect/designer* and *ETL developer*. In smaller organizations some individuals can have responsibilities from multiple roles, so this list should be only taken as a general guideline (*ibid, p.409*). And finally, when all the required stakeholders are involved, the DW project plan must be developed and then maintained, which also helps identify how and when the DPPs that cannot be implemented with technical means, should be embedded in the strategy, policies and processes embedded in the project.

3.6 Summary

This chapter outlined the methods and design strategies used for preparing the base for analysis and application of the proposed frameworks and best practices involved in the integration of DPPs with the design process of a DW system. To make sure that the right methodology is selected for this study, literature was reviewed on the topics related to research approaches in IS and business disciplines, and research design science was found as the most appropriate technique for this study.

Relevant privacy frameworks were combined to propose the finalized recommendations (*Table 6*) and categorization of technical and organizational measures that must be implemented to ensure development of GDPR-compliant system. Some of the core elements of DW architecture were reviewed, as these were important to understand in the context of this study. It was decided that Kimball's architecture is more appropriate for designing privacy, and to ensure consistency, recommendations of Kimball and Ross (*2013*) regarding project planning and development are also followed in this research.

Many studies from the reviewed literature focus on the perspective and concerns of end users, but it is hard to take research methodologies from those studies and apply them to questions that cover the topic from organizational standpoint. Other methodologies alone were not able to provide practical ways of answering the research question and achieving the expected results for this study; however, some could significantly complement the chosen approach – interviews or surveys could be used to identify whether proposed solution is feasible. Though this would have significantly widened the scope of the research and would not fit into requirements and limitations. With additional time and resources, the proposed solution could be further tested using other methods and techniques.

4 Framework Application and Analysis

4.1. Use Case Description

Following recommendations reviewed and methodology defined in previous chapters, this part of the dissertation will be dedicated to testing application of DPPs by outlining the process and then analyzing the results. Both technical and organizational measures will be discussed, and the business model to test the solution is an e-commerce website.

Let us say the example company is selling goods online to customers located in several countries in the EU. Users register on the website using their email address or via social media profile. Information about active customers or customers who deactivated their accounts is stored in the operational database that is accessed by application used by customer service teams as this information is required to answer customers' support queries. Users who access the website can browse it freely and see which products are available for sale, however if they would like to make a purchase – they are required to login to ensure that the order is tracked accordingly, linked to the right user and refund can be processed appropriately in case of return order. User experience with the website when they make a purchase can be described using three-tier architecture pictured in *Figure 12*. All interaction between the customers and the portal from opening the webpage, navigating products page, adding products to the basket, logging into the account to placing an order and making payment are processed by web, application and database servers. Operational databases store all current information such as customers' accounts, orders, inventory, order shipping and payments. As the business has been slowly growing, the marketing team has identified a need to analyze historical information about sales and customers' behaviours and determine trends which would help the company to establish itself on the new markets and increase revenue by offering more relevant products to their customers. With the support of company's CEO, it was decided to start a project for development of decision support system that would capture the required data and transform it into valuable strategic information. After evaluation of cost and resources, it was concluded that the DW will be receiving source data from the operational databases and transform it for analysis.

4.2. Project Planning and Requirements Definition

As the project now has the support of company's CEO, it is time to move on to further planning. CEO is well familiar with GDPR requirements and needs to make sure that the new system is compliant with the Regulation, so he assigned the Program Manager to lead the project, involve relevant stakeholders and documented all processes accordingly. First task that was given to the Program Manager was to make sure that all the involved parties are trained on GDPR matters, can identify privacy-threatening issues and offer solutions to

eliminate them, which guarantees compliance with the "accountability" DPP at the early stage of the project development. *Figure 16* represents Kimball's DW lifecycle diagram with privacy principles fit into the relevant parts of the DW project, as further outlined.

After ensuring that all involved parties are trained appropriately, the Program Manager sets up schedule, project roadmap and agrees on everyone's roles and responsibilities in the course of several meetings. One of the requirements for every member of the project group is that at every stage of the project development everyone is responsible for analyzing and capturing any privacy concerns that they can identify within the scope of their proficiency. For example, if business users decide that they need to be able to pull reports which would contain email addresses of the customer, then the business analyst might object to this request stating that providing report with this type of PII data will not benefit to the decision making process, but rather expose the users for potential disclosure. This ensures compliance with the "purpose limitation" DPP.

When roles and responsibilities are set, the team starts working towards defining business requirements. The business analyst, after getting back to the concerned departments and clarifying the needs, comes back with the list of potential questions that needs to be answered (as example – the list of questions is specified in *Appendix 4*). During the course of few meetings the project team evaluates the justification behind each question and finds out that around 10% of questions interfere with users' privacy and there is no legitimate purpose to analyze such sets of data. For example, "*List sales by product groups, ordered by IP address*", - since IP addresses can be used to directly identify a specific individual, there are more risks associated with pulling such a report than benefits. It was decided that the question can be rephrased to fit the purpose, but at the same time preserve the privacy of the individual user as much as possible: "*List sales by product groups, ordered by country/city*". Declining requests to query PII data that may be avoided during business requirements definition phase ensures compliance with the "data minimization" and "purpose limitation" DPPs.

After requirements are collected, next starts the process of designing and architecting the system. During this phase, responsible parties must ensure that relevant privacy preserving techniques are documented and applied in the design of the DW. This stage is crucial for setting up the relevant specifications for embedding "storage limitation", "integrity", "accuracy" and "data erasure" DPPs into the DW system: policies are created regarding data retention and archival terms, verification mechanisms are selected on how to identify that accurate and complete data is loaded from the operational database and relevant data protection techniques are chosen (encryption, anonymization, physical security, access restrictions and control, data separation and aggregation, logging and monitoring).

It is important to note that Kimball's DW design process resonates with the agile methodologies in software development. It means that previous task does not have to be completed before the next task of the workflow starts. If at any stage of the project any of the stakeholders would have any concerns, they can address the issues in an iterative manner while not breaking the whole workflow. However, when development of the system is finalized and everyone agrees on the launch date, it is important to inform customers that their data may be used in new ways. Before and during deployment, legal team can be involved to update the terms and conditions if required, and marketing team should be made responsible for sending relevant message to customers that their data may be now used for analytics (e.g., to improve purchasing experience, offer more relevant goods and services, expand in new geographies and increase the product base). At this stage, the "lawfulness, fairness and transparency" DPP should be covered.

After the new system is launched, it must be appropriately maintained – bugs fixed, features implemented, continuous operation of the system safeguarded. Since customers may contact the business with inquiries regarding which data about them is available to the company, "right of access" principle must be observed during DW maintenance. In addition to this, even though appropriate protection measures might have been implemented in the system, there is still no guarantee that this will protect the company and its customers from data breaches. Appropriate policies and processes must be established for the potential data disclosure event, whilst ensuring "confidentiality" principle.

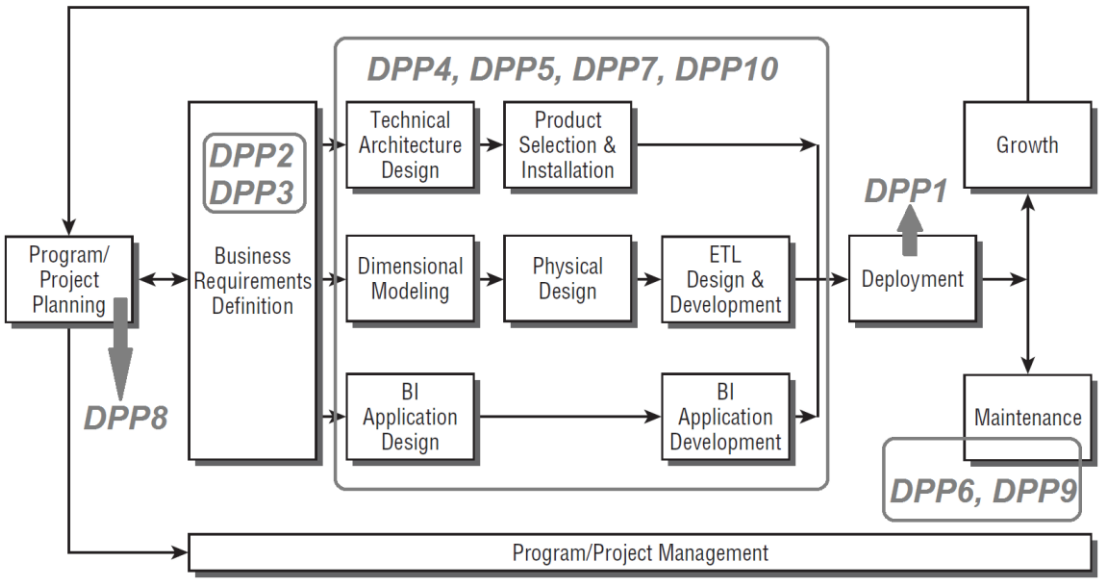


Figure 16: Mapping DPPs outlined in Table 5 with Kimball's DW lifecycle

Should there be a requirement to expand and grow the system, then it would trigger the process from the start considering the new scope. All the privacy principles would be expected to be followed respectively throughout the whole process.

4.3. Design and Development

When everyone is comfortable with their roles and responsibilities and after the business requirements have been set, the design and development phase of the project begins. It is not in the scope of this study to build the actual DW system, but rather outline the process and point out which privacy strategies can be applied at which stage of the SDLC. There is no need to develop schemas from scratch – there are resources available that recommend best practices and sample schemas for most business models. For example, Song and LeVan-Shultz (*1999*) developed a star schema for e-commerce sales, and it will be used as a base in this example (*Appendix 5*). As mentioned in the previous chapter, both star schema and OLAP cubes should be used in the architecture, as this ensures better security and access control to data of different levels of granularity. For example, each specific order will be stored in the star schema, but over time older data will be aggregated into higher level of granularity and loaded into OLAP cube, depending on the lifetime policy. During the design phase it is time to implement privacy strategies, and Hoepman's taxonomy is used for this purpose. While it is not in scope of this research to learn detailed technical characteristics of each privacy strategy or create new innovative methods of protecting privacy with technology, some examples are still provided. The ultimate goal of this modelling exercise is to find where and how privacy strategies fit in the design process.

Since the project team finalized a list of questions, for which the business needs answers, and defined which of them might potentially require collection or processing of PII data, it is now worth exploring one example of "partially" compliant question (defined in the previous chapter). Email addresses fall under the category of PII data as they can be used to identify an individual. This means that any analytical questions that might require tracking of customers' email addresses or activities that resulted from an event where an email was sent to a customer, must be approached with caution. One of such questions was: "How many people immediately "unsubscribe" when sent an e-mail notice?". One of the requirements for being compliant with GDPR is that companies cannot use collected personal data for purposes other than the original. Most companies store their customers' email addresses, and in most cases, there are legitimate reasons for doing so. Such reasons are typically addressed in the User Agreement (in other way known as "Terms and Conditions" of using the service). Some of the legitimate use cases for keeping customer's email address in the database could be registering an account, sending notifications about the status of their order, verifying account ownership or restoring access to account. These use cases generally benefit the customer directly. However, in case company decides to send promotional information to the customer's email, while the individual may have not agreed to receive such information, - not only could it be considered as a privacy violating activity, but depending on the context of promotion it could raise additional questions such

as which data was analyzed to make decisions about the promotion, where the data was coming from, where it is being stored, who has access and control over this data and how else it is being used. By using personal data for illegitimate purposes, not only companies risk their reputation, but this also makes them non-compliant with appropriate regulations. The following example describes one of the justifications to use customers' email addresses to answer an analytical question and shows how *data separation* privacy strategy may help with this task while keeping personal data protected.

The proposed sample online store needs to save customer's email address as a token that is used to login to their account. Email address plays an important role in terms of account security and protection since the owner of the email address in question is considered the owner of the account. If the business decided to quantify this information (e.g., find out how many customers use @gmail.com email address) to make any strategic decisions based on this data, the legitimacy of such task could be questioned as this was not the original purpose for which customers' email addresses were collected. If the organization wanted to answer the above question, then DW developers could separate, hide or anonymize this data, so that email address would not be a subject to data processing, and this information would not be visible to users with no valid business reasons to view it. Without DPbD philosophy in mind, the table containing the relevant information in the data warehouse would have each email recorded in plain format like presented in *Figure 17*:

email_address	email_subject	email_body	email_sent	unsubscribed	unsubscribe_timestamp	DW_timestamp
john.smith@gmail.com	Welcome!	Thanks for registering!	01.01.2019 0:00	FALSE	NULL	01.01.2019 0:00
john.smith@gmail.com	Welcome!	Thanks for registering!	01.01.2019 0:00	TRUE	02.01.2019 12:34	02.01.2019 12:34
sarah.doe@gmail.com	Welcome!	Thanks for registering!	05.01.2019 0:00	FALSE	NULL	05.01.2019 0:00
kate.white@gmail.com	Welcome!	Thanks for registering!	07.01.2019 0:00	FALSE	NULL	07.01.2019 0:00
kate.white@gmail.com	Welcome!	Thanks for registering!	07.01.2019 0:00	TRUE	08.01.2019 15:33	08.01.2019 15:33

Figure 17: Non-compliant example of storing email addresses in the DW

Each row would allow to track every "Welcome!" email sent to each new customer that registered on the website and whether customer clicked the "unsubscribe" link from that email. The record about the "unsubscribe" link provided in the body of the email would be stored in the same row, so one would know that customer unsubscribed after receiving this specific email. The table would also store appropriate timestamps, which would allow to track historical records – for example, how much time have passed since the time customer received the email until he unsubscribed. Having this data in one table would allow to answer the question: from the example provided in *Figure 17*, two clients have unsubscribed as soon as they received their "Welcome!" email. Even though this number alone does not help to identify the reason why this happened, and this data would need to be coupled with additional information to answer that question, having email address of the customer exposed to people who should not have access to this data, leaves the individuals' personal

information potentially vulnerable and open for disclosure. Example in *Figure 18* proposes a way to design the same table differently, with privacy in mind. Let us say that for each email notification (depending on the type of notification), there could be specific identifier generated to map that email to a unique number (ID). Instead of storing an actual email address directly in the table, only the ID would be visible.

link_id	email_subject	email_body	email_sent	unsubscribed	unsubscribe_timestamp	DW_timestamp
123123123	Welcome!	Thanks for registering!	01.01.2019 0:00	FALSE	NULL	01.01.2019 0:00
123123123	Welcome!	Thanks for registering!	01.01.2019 0:00	TRUE	02.01.2019 12:34	02.01.2019 12:34
2323232	Welcome!	Thanks for registering!	05.01.2019 0:00	FALSE	NULL	05.01.2019 0:00
543543543	Welcome!	Thanks for registering!	07.01.2019 0:00	FALSE	NULL	07.01.2019 0:00
543543543	Welcome!	Thanks for registering!	07.01.2019 0:00	TRUE	08.01.2019 15:33	08.01.2019 15:33

Figure 18: Compliant example of storing email addresses in the DW

It is still possible to answer the original question and determine how many customers unsubscribed after receiving their "Welcome!" email, however the table does not contain any sensitive PII data. Keeping records with mappings of "`link_id`" and related email address in order to track who and when unsubscribed from email notifications may be still required (security, integrity, transparency). Customer may come back and lodge a complaint that they did not opt out of notifications and that the company allegedly did it without their consent. In that case, a separate table with "`link_id`" to email address mappings should be stored, which would help to present this data back to the customer upon request.

link_id	email_address
123123123	john.smith@gmail.com
123123123	john.smith@gmail.com
2323232	sarah.doe@gmail.com
543543543	kate.white@gmail.com
543543543	kate.white@gmail.com

Figure 19: Mapping email address to its unique ID in a separate table

This information, if requested, may be presented to the customer with a message in their account specifying the timestamp when they unsubscribed. Alternatively, a limited number of people in the organization (presumably, customer service teams) might have access to this information and provide it back to the customer on demand. The team would not have access to the DW directly, but they would be provided with the relevant tools and applications, which in turn would pull this data from a DW source.

The above basic example described privacy techniques that can cover several DPPs. It was identified in the previous chapter that Hoepman's taxonomy covers all DPPs as far as this research is concerned, so following extensive analysis of the proposed privacy design strategies, it was identified where they fit in the Kimball's DW/BI architecture. *Figure 20* represents improved DW architecture with mapping of privacy strategies as they should be implemented in the design and development process of the system.

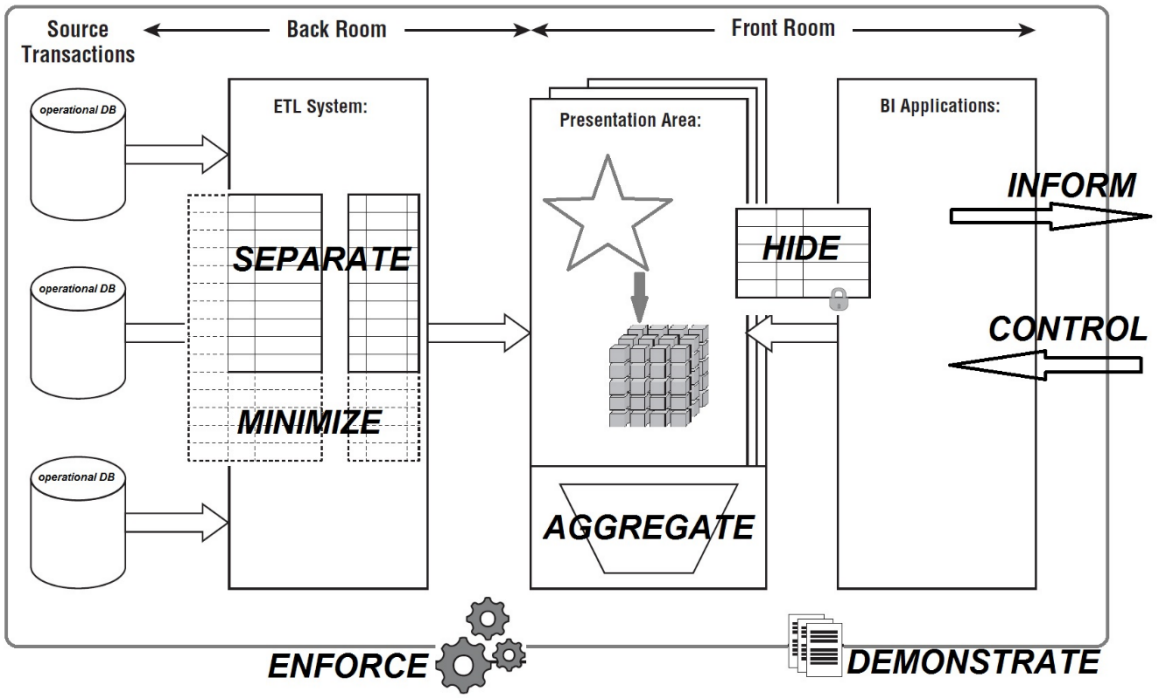


Figure 20: Application of Hoepman's privacy design strategies to Kimlall's DW/BI architecture (Hoepman et al, 2014; Kimball and Ross, 2013, p.19)

4.4. Analysis and Evaluation

To build a proposed framework, related literature was first reviewed on the relevant topics and some of the most suitable frameworks selected for further analysis. Then privacy strategies were mapped to the relevant DPPs and categorized based on their technical and organizational characteristics. Analysis of most common DW architectures was performed in order to choose the most suitable design that would resonate with the privacy strategies. While there are certain limitations to the proposed solution, overall architecture was accurately captured with the best practices and recommendations. Most difficulties that were encountered throughout the process were related to the lack of knowledge or experience with certain technical elements, but this was clearly defined in the scope of the study. One might argue that interpretation of the solution is subjective and requires further testing, however analysis and framework design activities were performed strictly within the limitations of this research. The goal was to outline the process how GDPR-compliant DW system can be implemented using available techniques, so it was tested on an example of e-commerce website, and finalized consolidated frameworks were presented in *Table 6* (data protection framework with technical and organizational measures recommendations), *Figure 16* (consideration of DPPs at all stages of DW project) and *Figure 20* (finalized framework of GDPR-compliant DW design).

5 Conclusions and Future Work

This chapter finalizes the research with conclusions and recommendations regarding future work in the areas discussed in this study.

5.1. Research Objectives: Summary of Findings

The goal of this study was to conduct extensive interdisciplinary research and investigate how technical and organizational measures can be coupled in systems development in order to build GDPR-compliant systems. When the GDPR was proposed, many researchers started investigating the Regulation from the perspective of users. A lot of studies represent the Regulation from the side where users have rights and organizations have obligations, which automatically assumes that there is no benefit for companies in becoming GDPR-compliant – it costs money, time and resources. However, many companies nowadays find innovative ways to do the right thing for society by applying appropriate measures to protect privacy and retain trust of customers. Determining optimal processes and recommendations for building systems with privacy in mind was the goal of this study. Several secondary objectives were set in the course of this research, and they were successfully fulfilled:

- Relevant and credible literature sources identified and reviewed
- Available privacy frameworks and design strategies analyzed
- Appropriate DW architecture selected for the research
- DW project phases and design process outlined using privacy frameworks

The primary research question was answered by combining multiple privacy frameworks with DW design strategies, and optimal recommendations were provided as a result of analysis and testing of the proposed solution on an example of an e-commerce website.

A few areas were identified where further research was needed, and some of the secondary questions were answered in the meantime.

The GDPR and DPbD principle

Even though the GDPR is not unique in its nature, the core privacy principles are the same across all similar regulations. While these principles propagate the idea of privacy as a basic human right, organizations struggle to understand how policies and regulations can be implemented using technical tools. The GDPR itself provides information regarding *what* should be done; however, it is pretty vague in regard to *how* it should be done. Nevertheless, since the GDPR is now mandatory by law, companies have no choice by to comply, so businesses are now looking for answers and recommendations regarding best practices on how GDPR compliance can be achieved. It is argued in this study that DPbD is one of the key principles of GDPR, and companies must design their systems with privacy in mind.

Data as an Asset

In the last decade, data has been called as "new oil" – it was considered that the more data companies have the higher the chance to win the competitive race, as data helps businesses make important strategic decisions. However, relevant privacy regulations have changed the game. Nowadays, the focus is made on trust and security. Customers state that they would not do business with the company who suffered from data breach and claim that companies do not take their customers' privacy seriously. Nevertheless, it appears that users easily transfer the responsibility for the security of their data on organizations while failing to take steps to secure their own data themselves. To ensure a win-win situation in these conditions, collaboration is required from both sides.

Data Analytics

Modern organizations collect, store, process and analyze a lot of data, and the amount of data generated in the world grows exponentially. However, some companies have no idea how much data they possess and what they are supposed to do with it. Transforming data in new ways and deriving valuable information from it is one of the main goals of business. There are various techniques available for organizations that help them derive value from data – BI applications, Data Warehouses, multiple data visualization applications and tools for big data analysis. It is important to understand the purpose and benefits of available tools and only work with the ones that fit the business goals.

Data Warehousing

There are few core DW architectures that are widely used by enterprises. In the course of this research, it was discovered that Kimball's architecture is more suitable for designing privacy. Some of the key concepts of DW design and architecture were discussed and compared, and the process of DW project development was reviewed and then tested on an example of an e-commerce website while applying privacy framework.

Overall, this dissertation fit into the scope of the study that was defined at the beginning of the project – there was no focus on technical details, but rather the issue was investigated from the organizational policy/process perspective.

5.2. Limitations of the Research

As any academic study that has strict deadlines, the main limitation was time. The period of this research was approximately six months from the time when research question worth investigating was identified until the chapters outlined for submission. It is worth noting that the original idea for this dissertation was different, but in the end, it served as a motivation for the research topic that was explored.

Another major limitation was related to the lack of knowledge and experience with the topic. Only general idea of the GDPR was known, mostly from the context of online blogs and articles covering the next data breach, and practical experience working with DW was limited. Some knowledge of data strategy in organizations was gained during the first year of study on the course. All of the DPbD and PbD techniques, strategies and considerations were unfamiliar, and thus required extensive research.

It can be argued that the research is subjective and was not tested in the field, so the proposed solution is open to interpretation and further additions. It was identified that there is no universal method on how to "hardcode" privacy, and even amongst known techniques, not everything is technically implementable. Therefore, another limitation of this research is that there is no right or wrong answer, so it is hard to evaluate the results objectively.

5.3. Suggestions for Future Work

There are many areas for future research since the topic is broad enough, and new interpretations, methods and techniques can be discovered over time. Some ideas around domains worth investigating in terms of privacy in DW systems can be: formalized privacy framework for DW program managers, defined roles of stakeholders involved in the DW project and their responsibilities in terms of preserving privacy in IS, best technical measures for privacy protection in DW systems (e.g., choosing suitable anonymization and encryption techniques, selecting appropriate access management tools and policies), the workflow for organizational privacy strategies or more detailed examples of specific privacy-preserving technical tools and techniques. Other frameworks and design techniques can be used as well, and one of the most feasible recommendations is that the solution proposed in this study can be tested using other methods such as interviews or surveys, and then optimized and improved after results evaluation.

5.4. Summary

Overall, the goals and objectives of this research were achieved, consolidated artifacts delivered as planned (*Table 6*, *Figure 16* and *Figure 20*), solution to the problem proposed and research question answered. All secondary areas and topics for research were discovered and sub-questions answered. The study stayed within the scope defined at the start of the research and fit into limitations and requirements. It may be concluded that the achieved results were in line with the expectations.

References

Blix, F., Elshekeil, S.A. and Laoyookhong, S., 2017. Data Protection by Design in Systems Development: From legal requirements to technical solutions. *The 12th International Conference for Internet Technology and Secured Transactions (ICITST-2017)*, December 2017, pp.98-103. 10.23919/ICITST.2017.8356355.

Breach Level Index, n.d. *The Breach Level Index,* [online] Available at: https://breachlevelindex.com/ [Accessed 24 April 2019]

BSI, (n.d.). Privacy matters: ISO/IEC 27552 whitepaper. *BSI Group,* [online] Available at: https://www.bsigroup.com/en-GB/iso-27552-privacy-information-management/ [Accessed 18 March 2019]

Cavoukian, A., 2009. *Privacy by Design: The 7 Foundational Principles,* [online] Available at: http://www.privacybydesign.ca [Accessed 20 January 2019]

Cavoukian, A., Taylor, S. and Abrams, M.E., 2010. Privacy by Design: essential for organizational accountability and strong business practices. *Identity in the Information Society,* [e-journal]. 3(2), August 2010, pp.405-413. 10.1007/s12394-010-0053-z.

Chalcraft, J., 2018. *Drawing Ethical Boundaries for Data Analytics*. Information Management Magazine. January/February 2018, pp. 18-22.

Chessell, M., 2014. Ethics for big data and analytics. *IBM Big Data & Analytics Hub,* [online] Available at: https://www.ibmbigdatahub.com/whitepaper/ethics-big-data-and-analytics [Accessed 10 March 2019]

Cho, Y.-S., Hoel, T. and Chen, W., 2015. Mapping a Privacy Framework to a Reference Model of Learning Analytics. *LACE Project,* [online] Available at: http://www.laceproject.eu/wp-content/uploads/2015/12/ep4la2016_paper_4.pdf [Accessed 18 March 2019]

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

ENISA, 2014. *Privacy and Data Protection by Design – from policy to engineering*. Heraklion: European Union Agency for Network and Information Security, December 2014. 10.2824/38623.

ENISA, 2019. *Guidance and gaps analysis for European standardization: privacy standards in the information security context*. European Union Agency for Network and Information Security (ENISA). 10.2824/698562.

EU GDPR.ORG, n.d. *EU GDPR Portal.* [online] Available at: https://eugdpr.org/ [Accessed 19 April 2019]

European Commission, 2019. *GDPR in numbers,* [infographic]. Available at: https://ec.europa.eu/commission/sites/beta-political/files/190125_gdpr_infographics_v4.pdf [Accessed 19 April 2019]

European Commission, n.d.(a). *Justice and Fundamental Rights,* [online] Commission and its priorities. Available at: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights_en [Accessed 19 April 2019]

European Commission, n.d.(b). *A new era for data protection in the EU: what changes after May 2018* [factsheet]. Available at: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf [Accessed 19 April 2019]

European Union, 2018. GDPR: new opportunities, new obligations. *Luxembourg: Publications Office of the European Union, 2018* [factsheet]. Available at: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-sme-obligations_en.pdf [Accessed 19 April 2019]

Fabian, B. and Goethling, T., 2015. Privacy-Preserving Data Warehousing. *International Journal of Business Intelligence and Data Mining*, [e-journal]. 10(4), September 2015, pp.297-336. 10.1504/IJBIDM.2015.072210

Galliers, R.D., 1995. A Manifesto for Information Management Research. *British Journal of Management*, [e-journal]. 6, December 1995, pp. S45-S52. 10.1111/j.1467-8551.1995.tb00137.x.

Gemalto, 2017. *Data Breaches and Customer Loyalty 2017 Infographic* [infographic]. Available at: https://safenet.gemalto.com/resources/data-protection/data-breaches-customer-loyalty-2017-infographic/ [Accessed 20 April 2019]

Greengard, S., 2018. Weighing the Impact of the GDPR. *Communications of the ACM*, [e-journal]. 61(11), October 2018, pp.16-18. 10.1145/3276744.

Hadar, I., Hasson, T., Ayalon, O., Toch, E., Birnhack, M.D., Sherman, S. and Balissa, A., 2017. Privacy by Designers: Software Developers' Privacy Mindset. *Empirical Software Engineering*, [e-journal]. 23(1), April 2017, pp.259-289. 10.2139/ssrn.2413498.

Hafiz, M., 2006. A collection of privacy design patterns. *In Proceedings of the 2006 conference on Pattern languages of programs (PLoP '06),* [e-journal]. ACM, New York, NY, USA, Article 7, October 2006. 10.1145/1415472.1415481.

Hevner, A., March, S.T., Park, J. and Ram, S., 2004. Design Science in Information Systems Research. *MIS Quarterly*, 28(1). March 2004, pp. 75-105.

Hoare, P., 2018. Over 3,200 the GDPR breaches logged. *Irish Examiner,* [online] Available at: https://www.irishexaminer.com/breakingnews/business/over-3200-gdpr-breaches-logged-894333.html [Accessed 24 April 2019]

Hoepman, J.-H., Cuppens-Boulahia, N., Cuppens, F., Jajodia, S., El Kalam, A.A. and Sans, T., 2014. *Privacy Design Strategies*. 29th IFIP International Information Security Conference (SEC), Jun 2014, Marrakech, Morocco. Springer, IFIP Advances in Information and Communication Technology, AICT-428, pp.446-459, 2014, ICT Systems Security and Privacy Protection. 10.1007/978-3-642-55415-5_38.

Hoepman, J.-H., 2018. *Privacy Design Strategies (The Little Blue Book)*, [monograph]. Nijmegen, Radboud University. [online] Available at: https://www.cs.ru.nl/~jhh/publications/pds-booklet.pdf [Accessed 6 April 2019]

ICO, n.d.(a). *Controllers and processors*. [online] Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/controllers-and-processors [Accessed 3 March 2019]

ICO, n.d.(b) *Guide to the General Data Protection Regulation (the GDPR)*. [online] Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/ [Accessed 3 March 2019]

Inmon, W.H., 2005. *Building the Data Warehouse*. 4th ed. Indianapolis, Indiana: Wiley Publishing, Inc.

Intersoft Consulting, n.d. Recital 78 Appropriate technical and organizational measures. *GDPR Recitals*, [online] Available at: https://gdpr-info.eu/recitals/no-78/ [Accessed 19 April 2019]

Jasmontaite, L., Kamara, I., Zanfir-Fortuna, G. and Leucci, S., 2018. Data Protection by Design and by Default: Framing Guiding Principles into Legal Obligations in the GDPR. *European Data Protection Law Review*, [e-journal]. 4(2), pp.168-189. 10.21552/edpl/2018/2/7.

Kimball, R. and Ross, M., 2013. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd ed. Indianapolis, Indiana: John Wiley & Sons, Inc.

Koops, B.J, and Leenes, R.E., 2014. Privacy regulation cannot be hardcoded. A critical comment on the 'privacy by design' provision in data-protection law. *International Review*

of Law, Computers & Technology, [e-journal]. 28(2), March 2014, pp.159-171. 10.1080/13600869.2013.801589.

Kroener, I. and Wright, D., 2014. A Strategy for Operationalizing Privacy by Design. *The Information Society,* [e-journal]. 30(5), October 2014, pp.355–365. 10.1080/01972243.2014.944730.

Kugler, L., 2018. The war over the value of personal data. *Communications of the ACM*, [e-journal]. 61(2), February 2018, pp.17-19. 10.1145/3171580.

Lapowsky, I., 2018. Facebook Exposed 87 Million Users to Cambridge Analytica. *Wired*, [online] 4 April. Available at: https://www.wired.com/story/facebook-exposed-87-million-users-to-cambridge-analytica/ [Accessed 19 April 2019]

Lord, N., 2018. What is the Data Protection Directive? The Predecessor to the GDPR. *Digital Guardian* [blog], September 2018. Available at: https://digitalguardian.com/blog/what-data-protection-directive-predecessor-gdpr [Accessed 20 April 2019]

Luján-Mora, S. and Trujillo, J., 2004. A Data Warehouse Engineering Process. *In: proceedings of the 3rd International Conference in Advances in Information Systems*, pp.14-23: Lecture Notes in Computer Science 3261, Izmir (Turkey), October 2004. 10.1007/978-3-540-30198-1_3.

McCallister, E., Grance, K. and Scarfone, K., 2010. *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*. Technical Report. NIST, Gaithersburg, MD, United States. SP 800-122.

McDougall, S., 2019. Adtech fact finding forum shows consensus on need for change. *ICO news, blogs and speeches*, [blog] 7 March. Available at: https://ico.org.uk/about-the-ico/news-and-events/blog-adtech-fact-finding-forum-shows-consensus-on-need-for-change/ [Accessed 23 March 2019]

Nagaty, K.A., 2010. E-Commerce Business Models: Part 1. *IGI Global*, pp.347-349. [online] Available at: http://www.irma-international.org/viewtitle/41196/

Narayanan, A. and Shmatikov, V., 2010. Myths and fallacies of "Personally Identifiable Information". *Communications of the ACM*, [e-journal]. 53(6), June 2010, pp.24-26. 10.1145/1743546.1743558.

Offermann, P., Levina, O., Schroenherr, M. and Bub, U., 2009. Outline of a design science research process. *In: proceedings of the 4th International Conference on Design*

Science Research in Information Systems and Technology, DESRIST 2009, Philadelphia, Pennsylvania, USA, May 7-8, 2009. 10.1145/1555619.1555629.

Parker, C.M., Wafula, E.N., Swatman, P.M.C and Swatman, P.A., 1994. Information Systems Research Methods: The Technology Transfer Problem. *In proceedings of: 5th Australasian Conference on Information Systems*, September 1994, pp.197-208.

Poniah, P., 2010. *Data Warehousing Fundamentals for IT Professionals*. 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

Rommetveit, K., Tanas, A. and Van Dijk, N., 2018. Data protection by design: promises and perils in crossing the Rubicon between law and engineering. *Forthcoming in: proceedings IFIP Summer School 2017 on Privacy and Identity Management*, May 2018. 10.13140/RG.2.2.24024.19202.

Sidgman, J. and Crompton, M., 2016. Valuing Personal Data to Foster Privacy: A Thought Experiment and Opportunities for Research. *Journal of Information Systems*, [e-journal]. 30(2), Summer 2016, pp.169-181. 10.2308/isys-51429.

Song, IY. and LeVan-Shultz, K., 1999. Data Warehouse Design for E-Commerce Environment. In: proceedings of ER Workshops 1999. Lecture Notes in Computer Science, vol 1727. Springer, Berlin, Heidelberg. 10.1007/3-540-48054-4_30.

Spiekermann, S. and Cranor, L.F., 2009. Engineering Privacy. *IEEE Transactions on Software Engineering*. 35(1), January/February 2009, pp.67-81.

Sweeney, L., 2002. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, [e-journal]. 10(5), October 2002, pp.557-570. DOI: 10.1142/S0218488502001648.

Symantec, 2018. How Common Is Identity Theft? (Updated 2018) The Latest Stats. *Symantec,* [online] Available at: https://www.lifelock.com/learn-identity-theft-resources-how-common-is-identity-theft.html [Accessed 27 April 2019]

Tene, O. and Polonetsky, J., 2013. Big Data for All: Privacy and User Control in the Age of Analytics. *Northwestern Journal of Technology and Intellectual Property*. 11(5), April 2013, pp.239-274.

The Economist, 2019a. Europe's the GDPR offers privacy groups new ways to challenge adtech: Whose business is your data? *The Economist,* [online] Available at: https://www.economist.com/briefing/2019/03/23/europes-gdpr-offers-privacy-groups-new-ways-to-challenge-adtech [Accessed 24 March 2019]

The Economist, 2019b. Big tech faces competition and privacy concerns in Brussels. *The Economist,* [online] Available at: https://www.economist.com/briefing/2019/03/23/big-tech-faces-competition-and-privacy-concerns-in-brussels [Accessed 24 March 2019]

Truex, D.P., Holmstroem, J. and Keil, M., 2006. Theorizing in information systems research: A reflexive analysis of the adaptation of theory in information systems research. *Journal of the Association for Information Systems,* [e-journal]. 7(12), December 2006, pp.797-821. 10.17705/1jais.00109.

Turner, N. and Burbank D., 2016. Brexit & the day-to-day role of the CIO: a data management & governance perspective. *Enterprise Management 360º*, [online] Available at: https://globaldatastrategy.files.wordpress.com/2015/09/brexit-gds-em360-q42016.pdf [Accessed 3 February 2019]

Vaishnavi, V., Kuechler, B., Petter, S. and De Leoz, G., 2004/17. *Design science research in information systems.* [online] Available at: http://www.desrist.org/design-research-in-information-systems/ [Accessed 27 January 2019]

Van Dijk, N., Rommetveit, K., Tanas, A. and Raab, C., 2018. Right Engineering? The redesign of privacy and personal data protection. *International Review of Law, Computers & Technology,* [e-journal]. 32(2-3), April 2018, pp.230-256. 10.1080/13600869.2018.1457002

Van Hoof, O., 2017. GDPR Forcing Organizations to View Data Strategically. *Transforming Data with Intelligence,* [online] November 2017. Available at: https://tdwi.org/articles/2017/11/14/twt-all-gdpr-forcing-organizations-view-data-strategically.aspx [Accessed 20 April 2019]

Bibliography

Brendel, A.B., Zapadka, P. and Kolbe, L.M., 2018. Design Science Research in Green IS: Analyzing the Past to Guide Future Research. *In: proceedings of the European Conference on Information Systems.*

Danezis, G. and Guerses, S., 2010. A critical review of 10 years of Privacy Technology. *Microsoft Research, Cambridge, U.K*

Davis, C., 2019. GDPR and CRM: How to Manage Customer Data in 2019. *SuperOffice*, [blog]. Available at: https://www.superoffice.com/blog/gdpr-crm/ [Accessed 27 April 2019]

ENISA, 2015. *Privacy by design in big data: an overview of privacy enhancing technologies in the era of big data analytics*. European Union Agency for Network and Information Security (ENISA). 10.2824/641480.

Galliers, R.D., 1991. Choosing Appropriate Information Systems Research Approaches: A Revised Taxonomy. *Information Systems Research: Contemporary Approaches and Emergent Traditions. Elsevier Science Publishers B.V.* (North-Holland), pp.327-345.

Gonzalez, R. and Dahanayake, A., 2007. *A Concept Map of Information Systems Research Approaches*. 18th Annual IRMA International Conference. Managing Worldwide Operations & Communications with Information Technology, May 2007, pp.845-848.

Hornung, G., 2013. Regulating privacy enhancing technologies: seizing the opportunity of the future European Data Protection Framework. *Innovation: The European Journal of Social Science Research,* [e-journal]. 26(1-2), pp.181-196. 10.1080/13511610.2013.723381.

Irwin, L., 2018. The GDPR: How the right to be forgotten affects backups. *IT Governance European Blog*, [blog]. Available at: https://www.itgovernance.eu/blog/en/the-gdpr-how-the-right-to-be-forgotten-affects-backups [Accessed 27 April 2019]

McCreanor, N., 2018. How to maintain GDPR-compliant databases. *IT Governance European Blog*, [blog]. Available at: https://www.itgovernance.eu/blog/en/how-to-maintain-gdpr-compliant-databases [Accessed 27 April 2019]

Peffers, K., Tuunanen, T., Gengler, C.E., Rossi, M., Hui, W., Virtanen, V. and Bragge, J., 2006. *The design science research process: A model for producing and presenting information systems research*. DESRIST 2006, February 24-25, 2006, Claremont, CA.

Pfitzmann, A. and Hansen, M., 2008. *Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for*

Terminology. [online] Available at http://dud.inf.tu-dresden.de/Anon_Terminology.shtml [Accessed 13 April 2019]

Solove, D.J., 2008. Understanding Provacy (Chapter One). *Harvard University Press*, Cambridge, Massachusetts; London, England 2008. Available at: http://ssrn.com/abstract=1127888 [Accessed 20 April 2019]

Stackify, n.d. *What is N-Tier Architecture? How It Works, Examples, Tutorials, and More*. [online] Available at: https://stackify.com/n-tier-architecture/ [Accessed 27 April 2019]

Teece, D., Peteraf, M. and Leih, S., 2016. Organizational Agility: Risk, Uncertainty, And Strategy in The Innovation Economy. *California Management Review*, [e-journal]. 58(4), Summer 2016, pp.13–35. 10.1525/cmr.2016.58.4.13.

The Ecomonist, 2019. Why big tech should fear Europe. *The Ecomonist*, [online]. Available at: https://www.economist.com/leaders/2019/03/23/why-big-tech-should-fear-europe [Accesed 24 March 2019]

Van Rest, J., Boonstra, D., Everts, M., Van Rijn, M. and Van Paassen, R., 2014. *Designing Privacy-by-Design.* B. Preneel and D. Ikonomou (Eds.): APF 2012, Privacy Technologies and Policy, pp.55–72.

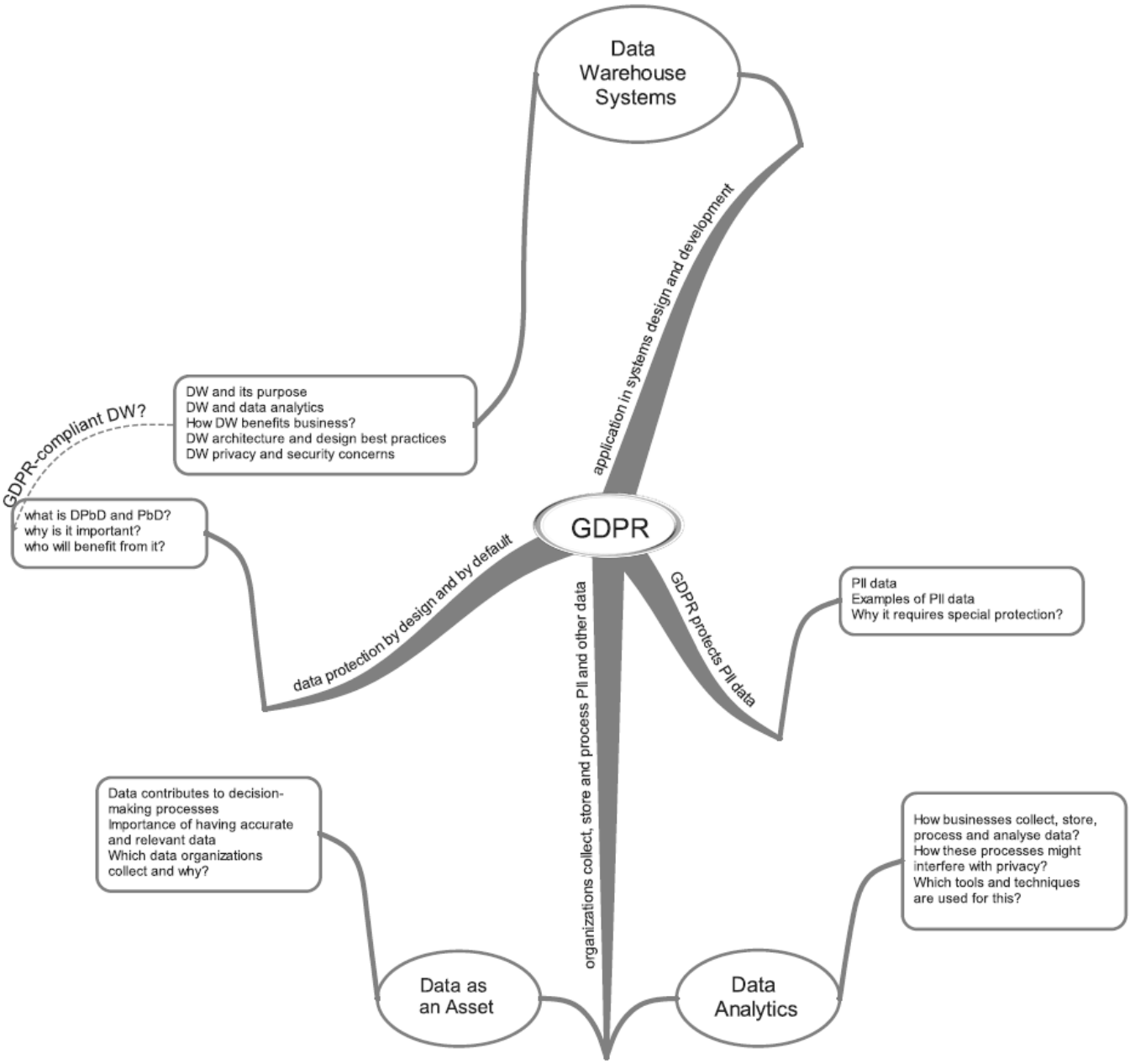
Welch, R., n.d.(a). *GDPR's Impact on BI (Part 1 in a Series)*. [online] Available at: https://tdwi.org/articles/2018/06/04/biz-all-gdpr-impact-on-bi-1.aspx

Welch, R., n.d.(b). *The 6 Pillars of the GDPR (Part 2 in a Series)*. [online] Available at: https://tdwi.org/articles/2018/06/05/biz-all-6-pillars-of-gdpr-2.aspx

Welch, R., n.d.(b). *GDPR and Tokenizing Data (Part 3 in a Series)*. [online]. Available at: https://tdwi.org/articles/2018/06/06/biz-all-gdpr-and-tokenizing-data-3.aspx

Appendices

Appendix 1 – Mind map of research areas and their relation



Appendix 2 – Article 25 of the GDPR: Data protection by design and by default

Article 25

Data protection by design and by default

1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

3. An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this Article.

Appendix 3 – Article 5 of the GDPR: Principles relating to processing of personal data

Article 5

Principles relating to processing of personal data

1. Personal data shall be:

(a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency');

(b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation');

(c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');

(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy');

(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation');

(f) processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality').

2. The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('accountability').

Appendix 4 – List of questions that DW can answer as per Song and LeVan-Shultz (*1999*), categorized and tested against DPPs

Category and question to be translated into OLAP query	Compliant?
Sales & Market Analysis	
· What is the purchase history/pattern for repeated users?	Caution
· What type of customer spends the most money?	No
· What type of payment options is most common? By size of purchase? By socioeconomic level?	Caution
· What is the demand for Top 5 x's based on the time of year and location?	Yes
· List sales by product groups, ordered by IP address.	No
· Compared to the same month last year, what are the lowest 10% items sold?	Yes
· Of multiple product orders is there any correlation between the purchases of any products?	Yes
· Establish a profile of what products are bought by what type of clients.	Caution
· How many different vendors are typically in the customer's market basket?	Caution
· How much does a particular vendor attract one socio-economic group?	No
· Since our last price schedule adjustment which products have improved, and which have deteriorated?	Yes
· Do repeat customers make similar product purchases (within general product category) or is there variation in the purchasing each time?	Caution
· What types of products do repeat customers most often purchase?	Caution
· For each vendor, what are the top three products offered that are most often purchased?	Yes
· What are the top 5 most profitable products by product category and demographic location?	Yes
· What are the top ten products that customers purchased in conjunction with product X?	Yes
· Which products are also purchased when one of the top 5 selling items is also purchased?	Yes
· What products have not been sold online since X days?	Yes
· In what zip codes do the highest number of sales occur?	No
· What day of the week do we do the most business by each product category?	Yes
· What is our average volume of business per product category per sales channel?	Yes
· What items are requested but not available and how often and why?	Yes
· What is the best sales month for each product?	Yes
· What is the average number of products per customer order purchased from the website?	Caution
· What is the average order total for customer orders purchased from the website?	Caution
· How well do new items sell in their first month?	Yes
· What season is the worst for each product category?	Yes
· What % of first-time visitors actually make a purchase?	Caution
· How many items are in the average order?	Yes
· What products attract the most return business?	Yes

· Is there a geographic correlation to with time of the year for the sales pattern of a certain product?	Yes
· Which customers who have previously ordered by phone are now using the web site(s)?	Caution
· Of the customers who access the e-commerce site(s) and don't make a purchase, how many call and order via the phone?	Caution
· Are new product offerings being introduced to established customers?	Caution
· Based on history and known product plans, what are realistic, achievable targets for each product, time period and sales channel?	Yes
· What are the sales to plan percentage variation for this year? What are the planning discrepancies?	Yes
· Have some new products failed to achieve sales goal? And should they be withdrawn from online catalog?	Yes
· Are we on target to achieve the month-end, quarter-end or year-end sales goals, by product or by region?	Yes
Returns	
· How often was product X returned?	Yes
· How often did a customer request a refund and how often did they request an exchange for another product?	Caution
· What are the top 5 products which have been returned by customers after purchasing?	Yes
· Do customers who complain or return items make future purchases?	Caution
· Do certain customers repeatedly return items?	Caution
Website design & navigation analysis	
· At what time of day does the peak traffic occur?	Yes
· At what time of day does the most purchase traffic occur?	Yes
· Which types of navigation patterns result in the most sales?	Caution
· How often are purchasers looking at detailed product information by vendor types?	Yes
· What are the top ten most visited pages? (Per day, weekends, months, seasons)	Yes
· How much time is spent on pages with banners and without banners?	Yes
· How does a non-purchase correlate to web site navigation?	Caution
· Which vendors have the most hits?	Yes
· How often are comparisons asked for?	Yes
· Based on website page hits during a navigation path what products are inquired about the most but seldom purchased during a visit to our website?	Caution
· Do products with pictures and extended descriptions sell better than those without pictures?	Yes
· Where are high-spending customers surfing to our website from?	Caution
· How often do customers arrive at the website from their ISP's home page?	No
· How often do customers arrive at the website from a site containing an ad banner?	Caution
· How often do customers make a purchase when arriving from a website containing an ad banner?	Caution
· How often do customers arrive at the website from links contained in e-mail notification?	Caution
· Do most customers use the search engine or just browse the site?	Caution
· Do items highlighted on the main page sell better?	Yes

· What % of customers who leave items in shopping basket return later and purchase them?	Caution
· How does Internet traffic bandwidth affect the number of clients?	No
· What is the most popular search engine through purchaser's access?	Yes
· What are the top complaints about the web site(s)?	Yes
· What groups of customers find the web site(s) hardest to use?	Caution
· Make recommendations for future purchases to the client, based on what the client purchased in the past.	Caution
Customer service	
· What are the top 5 complaints about the products or services?	Yes
· Does e-mail notification of new products or price reductions to regular customers increase sales?	Caution
· How many people immediately "unsubscribe" when sent an e-mail notice?	Caution
· Did sales decrease after requiring users to register?	Caution
Warehouse/Inventory	
· Which locations provide a cost-effective restock to which locations?	Yes
· What is the average back-order time, i.e. the time, when a product is out of stock, from when a customer orders the item until it is back in stock and shipped?	Yes
· Do we have adequate inventory for a particular product to meet anticipate demand?	Yes
Promotions	
· After 10% discount promotion, what is the increase of sales for the products?	Yes
· To what extent did a promotion of a product effect sales of that product?	Yes
· Do sales incentives like "limited time offer" increase sales?	Yes
· Do discounts based on multiple purchases of an item increase sales?	Yes
· Do specials offered to best customers' result in increased sales?	Caution
· Is there a correlation between promotions and sales growth?	Yes
· Are some sales group achieving their monthly or quarterly targets by excessive discounting?	Yes
· What average discounts are being given for different products or channels?	Yes
· Is our advertising budget properly allocated? Do we see a rise in sales for products and in areas where we run campaigns? How much is the rise?	Yes
Shippings	
· Is there a change in delivery type at different times of the year, i.e. preceding major holidays?	Yes
· What is the average time from ordering date to shipping date? Does this vary by product?	Yes
· What types of delivery options are requested per each category and region?	Yes

Appendix 5 – Base Star Schema for e-commerce sales as per Song and LeVan-Shultz (*1999*)