KU LEUVEN

FACULTY OF ECONOMICS
AND BUSINESS
CAMPUS BRUSSELS

Internally Organised Master's Thesis

Development of a Capability Maturity Model for Big Data Governance

Evaluation in the Belgian Financial Sector

Andra-Raluca MERTILOS

Master's Thesis Submitted for the Degree of
Master in Business Administration
Graduation Subject: Business Information Management

Supervisor: Yves WAUTELET

Academic Year: 2014–2015

Defended in: June 2015

FACULTY OF ECONOMICS AND BUSINESS - CAMPUS BRUSSELS
WARMOESBERG 26 – B 1000 BRUSSELS BELGIUM

Table of contents

List of figures

List of tables

List of appendices

List of abbreviations used

ITGI	Information Technology Governance Institute
CMM	Capability maturity models
CMMI	Capability maturity model integration
DFA	Dodd-Frank Act
TQM	Total Quality Management

1. INTRODUCTION

A "*normal*" conversation between the front and back-office of any organization is a juggling of "*blames*" for what seems to work inefficiently. The back-office complains that the processes followed by the front office generate bottlenecks by failing to reflect how the business actually operates. On the other hand, the front office argues that the existing processes are designed to resolve potential shortfalls in business operations but that the back-office fails to comply with them. The same back-and-forth holds for data governance programs as well.

Data has become a major asset in today's business landscape: data about customers, operations, clients, suppliers or creditors occupies decision-makers agendas on a daily basis by driving evaluations and resolutions with the potential to impact a company at every level. This data being valid and accurate is then of central significance in determining the weight of outcomes and the influence they generate. Nowadays this becomes ever more complex as new data sources bring about challenges in terms of volume, variety or plurality, just to name a few. The "*Big data phenomenon*" has been characterized as one of the most discussed topics in research and practice with more than 70% of all ranked papers on this subject having been published only in the last two years (Buhl, Röglinger, Moser, Heidemann, 2013).

Bahjat El-Darwiche, Koch, Tohme, Shehadi and Meer (2014) point out that a common misconception when discussing about Big Data is that it revolves around complicated technologies which discourage companies from embarking on such initiatives. While we acknowledge that this has mostly been the case, the main driving force of success of any big data project is that organizations need to reshape the way decision-making is enforced, basing it more on clear data insights rather than just pure intuition. Big data projects will provide the promised results if they are built on the foundations of an environment which already fosters a data-driven culture and mindset. Big data is indeed not a magical fix for any data problems an organization might have. What it offers is the possibility to expand the decision-making scope by recognizing the multitude of angles of approach when addressing a business task. This ensures that both internal and external policymakers have all the information at hand to "*craft*" valid decisions. Otherwise put, that there exists a data governance framework in place to build upon.

1.2 Research context

Tamasauska, Liutvinavicius, Sakalauskas and Kriksciuniene (2013) characterize the data currently being used by the financial institutions as meeting all the requirements for big data (pp.36): "*massive, temporarily ordered, fast changing, potentially infinite*". According to them, successfully utilizing big data has the potential to bring about the necessary transformations in the banking sector (pp.36): "*create a customer-focused enterprise*", "*optimize enterprise risk management*", "*increase flexibility and streamline operations*". A few banks in the Benelux area are just embarking on big data initiatives such as the ING Group (Finance Lab, 2014) and KBC Belgium (Van Leemputten, 2014) and this novelty makes it difficult to build a big data governance program to suit current project needs as these needs in themselves are not yet properly documented or understood. Using the approach of De Haes & Van Grembergen (2005), what is needed in such cases is to draw on existing data governance structures and design a *capability maturity model* (CMM) which can steer projects in the right direction based on their own capabilities and needs. The levels of maturity for big data governance need to be synchronized with the needs of the organization. This can done by following a staged approach otherwise investing in complex Hadoop clusters will prove useless if we have no understanding of their purpose. The motivation for choosing a CMM to assess big data approaches can best be summarized by a quote of O'Regan (2011, pp.45): "*It (…) provides a roadmap for an organization to get from where it is today to a higher level of maturity*"

In the light of these insights, we have built our research around the following central research question:

What are the key process areas, common features, key practices and goals for each of the 5 levels of a capability maturity model regarding Big Data Governance practices in the Belgian Financial Sector?

1.3 Research process

In order to successfully answer this question, the central research question has been devised in the following research objectives:

1. Conduct a literature review to clarify the definitions and theories behind the following concepts :
2. Identify main elements of a capability maturity model and explain their structure:
3. Identify the process areas of existing data governance models by conducting a literature review of existing big data and data governance models.
4. Identify the most common big data dimensions.

5. Describe and characterize the specific characteristic of the financial sector in terms of financial data records and data collection practices.

6. Analyse and identify the most important data governance process areas as mentioned in the Basel III framework for risk reporting.

7. Map the model identified at point 3 with the model identified at point 5 and evaluate their fit.

8. Test the model at point 7 by evaluating it in the banking sector via qualitative interviews with subject matter experts and/or key banking representatives.

9. Draw final conclusion and recommendations.

Chapter 2 introduces the reader to the definition of governance and IT governance as we conduct a literature review to identify the components and mechanisms of what constitutes an IT governance framework by looking at how omnipresent IT has become in an enterprise, what role environmental contingencies play in shaping decision-making structures and fields as well as the scope and functions of these elements. Further, we present a commonly used IT governance framework by outlining its structures, processes and relational mechanisms.

Chapter 3 presents the concept of capability maturity models (CMM's) by analysing maturity, performance and capability as well as the origins of the first CMM's. We will look at how capability maturity models are structured and what are the different key process areas, levels of maturity, common goals and key practices defining their use.

Chapter 4 advances the concepts and theories behind data governance by explaining the differences between data and information as well as between governance and management. Borrowed concepts from IT governance such as the contingency theory and classifications on structures, operations and relations will help in explaining the processes, responsibilities and decision-making rights which will constitute the building blocks of our data governance model.

Chapter 5 defines and explains big data concepts and features by proposing a dimensions model, explaining the difference between traditional and big data as well as dealing with the plurality of data sources, technologies and architectures. It then looks at how big data projects are financially valuated, what are the sensible aspects associated to it and the potential use cases which can be derived from using big data.

Chapter 6 presents a view on the financial sector by focusing on practices related to data collection, aggregation and governance of financial records. During this chapter we are also acquainted with the specificities of big data governance programs. Further on, it brings together previous chapters by integrating and linking concepts and principles,

data and big data governance together in one capability maturity model fit to map against the current Basel III framework for risk data reporting. The section will also present the initial results of our short empirical tests.

The final chapter will present our conclusions and recommendations.

2. IT GOVERNANCE

2.1 About this chapter

As corporate governance begins to rely more and more on IT capabilities and resources to ensure business continuity, the necessity to understand how governance related concepts apply to IT domains in terms of policies, principles, strategies and guideline has risen. The subject of IT governance has been extensively researched in scientific literature with structures, processes and relational mechanisms identified, documented and categorized however without a homogeneous definition yet to encompass the major concepts behind such a framework. Reaching a common definition, along with the identified structures and topologies will help pinpoint the concepts and structures which help in describing, characterizing, designing and building an IT governance framework capable of ensuring a fusion between business and IT. Such a model should remain firmly grounded in a broader corporate governance context and build upon developing its elements with respect to the common objectives and goals as defined at organization-level. The following chapter will focus on defining and positioning IT in the enterprise as well as identifying and describing its specificities and characteristics, as a governance component and as a function. The last part will present the elements of the most common IT governance framework identified in our literature studies and explain its components and elements.

2.2 Governance and IT governance

This section defines the concepts of governance and IT governance.

2.2.1 Governance definition

Governance as a concept is conceptualized by using the agency theory (De Abreu Faria, Macada & Kumar, 2013) which is widely used in organizational studies to explain the relationship between a principal and an agent with regards to matters of control, risk, monitoring, rules, alignment and structure. Weber, Otto and Osterle (2009) define governance as (pp. 4:3) "*the way the organization goes about ensuring that strategies are set, monitored and achieved*". Datskovsky (2010, pp.158) defines governance as "*the set of processes, customs, policies, controls, regulations, and institutions that affect the way a corporation is directed, administered, or controlled"*. He emphasizes that, in an enterprise, different sources offer recommendations for policies and principles corresponding to different departments and parts of the organization: the company has

to translate these into an overall strategy and concrete guidelines for each organizational domain.

2.2.2 IT governance definition

According to Ploder and Fink (2008), issues of corporate governance have become more and more aligned to the IT needs and this has given rise to a new research field, namely *IT governance*. A lot of authors have focused on trying to provide a common definition for IT governance (Lewis & Millar, 2009; Simonsson & Ekstedt, 2006; Webb, Pollard & Ridley, 2006) and while no homogenous definition exists, most research has been targeted to analyzing and compiling existing definitions in literature and deriving potential components of an IT governance framework.

Peterson is mentioned by Lewis and Millar (2009) as defining IT governance in terms of decision rights and accountabilities regarding the desirable behavior in the use of IT. Another common definition mentioned in literature is the one mentioned by the Information Technology Governance Institute (ITGI)(Lewis & Millar, 2009, pp.2; Nassiri, Ghayekhloo & Shabgahi, 2009; Ploder, 2008): "*a structure of relationships and processes to direct and control the enterprise in order to achieve the enterprise's goals by adding value while balancing risk versus return over IT and it's processes*".

Simonsson and Ekstedt (2006, pp.20) propose a definition based on a compilation of approximately 60 scientific articles about IT governance, i.e.: "*decision-making upon certain assets, i.e. the hardware and software used, the processes employed, the personnel, and the strategic IT goals of the enterprise*". Webb et al. (2006) apply a content analysis approach to 12 common definitions (aggregated via a literature review) based on the number of occurrences of a number of elements identified as defining IT governance. The proposed definition is (pp.6): "[…] *the strategic alignment of IT with the business such that maximum business value is achieved through the development and maintenance of effective IT control and accountability, performance management and risk management*". Another common definition is the one given by Van Grembergen, De Haes and Guldentops (2004) as IT governance being the "*organizational capacity exercised by the Board, executive management and IT management to control the formulation and implementation of IT strategy and in this way, ensure the fusion of business and IT*".

As most definitions mention control structures, decision-making and strategic alignment of IT with the corporate objectives in contexts such as performance and risk management, it is important to further investigate how these constructs interact and shape up to building IT governance domains.

2.3 Components and mechanisms

This section describes how IT is positioned in an enterprise by analyzing its environmental contingencies as well as the scope and functions of decision-making structures and fields governing IT.

2.3.1 IT omnipresence in the enterprise

According to Ploder (2008), IT has moved from a support role to generating a competitive advantage and creating a sustainable value for the organization. Peterson (2003) mentions the "*pervasiveness*" of IT nowadays: decisions regarding IT can no longer be delegated or avoided by business managers like they were in the past. Heier, Borgman and Mileos (2009) also mention increasing IT omnipresence as one of the factors responsible for the augmented importance of IT in the strategy's success at corporate level along with compliance to regulations which request more transparency in business operations. They continue by presenting what they call the "*productivity paradox*" for investments made in IT: measuring IT budges does not provide measurable business value. Traditionally looking at IT investment budgets does not account for the increased complexity of offshoring and outsourcing IT preparations as well as for the surge in human and financial implications of IT investments. Such business value is derived from the proper implementation of governance applications and this implementation involves the undertaking of both quantitative and qualitative indicators for IT governance applications' success. Van Grembergen et al. (2004) position IT as a competitive advantage and stress its movement across the ladder from service provider to strategic partner.

2.3.2 Environmental contingencies

Lewis and Millar (2009) pointed out that the IT governance subject has, among others, also been influenced by such schools of thought as *methodological comprehensiveness* and *social interventions*. Ribbers, Peterson & Parker (2002) use *environmental contingencies* in their research to explain the relationship between IT governance and its outcome in the light of the schools of thought mentioned by Lewis and Millar (2009).

Figure 2.1 summarizes the relationship matrix between the environmental contingencies identified by Ribbers et al.(2002).

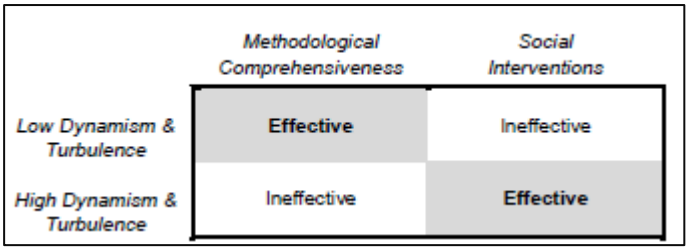

Figure 2.1 Environmental contingencies and effective decision-making (Ribbers et al., 2002, pp.2)

The 2 environmental contingencies identified by the authors are *dynamism* and *turbulence* and they influence IT governance outcomes in the following way: if low dynamism and low turbulence then IT decision-making is associated and perceived as highly methodological with low social interventions; on the contrary, if high dynamism and high turbulence then IT decision-making is less based on methodologies and more reliant on social interventions.

2.3.3 Decision-making structures

Research in the domain of IT governance has traditionally been concerned with the *decision-making structures* for IT control with orientations going from differentiation of IT decision-making structures towards integrating these structures in value maximization (Ribbers et al. , 2002). The authors attempt to characterize IT governance on the basis of an organizational model of problem identification and problem solution :

- *Problem identification* is concerned with scanning internal and external environments and identifying potential problems before they occur.
- *Problem solution* is concerned with implementing the necessary courses of action to stop these problems from occurring.

Simonsson and Ekstedt (2006) mention a difference in *levels of priority* regarding IT governance in literature and practice. More specifically, in literature, IT governance is more often than not defined as being the responsibility of Board of Directors and executive management in selecting and using key strategic relationships meant to obtain and reinforce IT competencies. They also mention IT governance as being at the crossroads between ensuring "*fusion*" between business and IT and for alignments between business, IT and the creation of value across the enterprise.

2.3.4 Scope of decision-making

Simonsson and Ekstedt (2006) divide the decision-making structures on which IT governance provides input on into dimensions because it helps with indicating the scope of the decision-making. The identified dimensions are:

- *Goals*: form of measurement of how well the objectives set will perform. These are mainly decisions regarding IT policies, corporate strategy relating to the use of IT, frameworks and objectives or roadmaps for kick-off

- *Processes*: implementation and management of the IT structures which will support operations. These include identifying and defining the relevant IT tasks, setting up procedures and nets for a good accomplishment of these tasks
- *People*: roles and accountabilities of the different participants. These decisions include defining structures of responsibility in the greater corporate context as well as on process-context, defining what each role does and what are the skills needed to fill in the different roles.
- *Technology* : physical assets such as hardware, software, facilities

2.3.5 Decision-making fields

Peter Weil (Lewis & Millar, 2009) draws upon the work of Peterson and identifies 3 types of governance mechanisms for IT decisions: *decision-making structures, alignment processes* and *communication approaches*. Weil and Ross (2004) had also analyzed institutional approaches on IT as well as decision-making structures from the point of view of domains, styles and mechanisms. They identified 5 key decision fields:

- *IT principles* position IT and its role in the business
- *IT architecture* addresses issues as data, application and infrastructure in the context of standardization
- *IT infrastructure* includes hardware and software common services
- *Business applications* needs are the liaison between IT and the accomplishment of the business strategy
- *IT investment and prioritization decisions* prioritize and rank projects according to resources and budgets

2.3.6 Functions

Nassiri et al. (2009) reference *function and value alignment* as being one of the key purposes of IT governance, along with *risk management, performance measurement and responsibility*. Webb et al. (2006) identify *strategic alignment, delivery of business value, risk and performance management* as being the elements of which most IT governance literature focuses on.

Van Grembergen et al. (2004) identify strategic alignment and business value as 2 important elements in IT governance. They define strategic alignment as (pp.7): "*the process and goal of achieving competitive advantage through developing and sustaining a symbiotic relationship between business and IT".* For Webb et al. (2006), business value is delivered by "*exploring opportunities and maximizing benefits* ".

Decision-making as the core of IT governance entails unlocking a number of steps such as having a solid understanding of what the underlying model enabling these decisions is and assessing the consequences which might be associated with it. Once this model has been created and understood, we can decide and plan how the decision has performed according to the established baseline by means of objectives and measurements. The scope of the decision-making process involves a short or long term vision and can be divided in strategic and tactical rulings on key elements composing an IT governance framework.

2.4 Building an IT governance framework

This section identifies and describes the components of an IT governance framework.

2.4.1 IT governance framework outline

Peterson is mentioned across literature as developing the first IT governance framework based on 3 components (Lewis & Millar, 2009; Nassiri et al., 2009): *structural capabilities, process capabilities* and *relational capabilities*. Put briefly, structural capabilities refer to people and organizational design of responsibility and functions, processes refer to the domains on which decision-making is done while relational capabilities refer to the means used to "*bridge the gap between business and IT*" (Nassiri et al., 2009, pp.218).

Van Grembergen et al.(2004) further develops the work of Peterson and Weil and builds a comprehensive IT governance framework composed of : *structures, processes and relational mechanisms.* Each of these elements will be presented in the following sections.

2.4.2 Structures

Structures refer to interactions between organizational levels and departments, as well as accountabilities and authority regarding policymaking and supervisory plans and strategies. Van Grembergen et al.(2004), distinguishes between 2 integration strategies at tactic and mechanisms level : tactics are more concerned to the positioning of programs authority at corporate level while mechanisms distil these policies in plans, rules, guidelines and tasks. The manner in which the 2 levels collaborate with each other depends on the IT organization structure and to how the level of authority regarding IT decision-making moves away or towards a corporate Information Systems strategy (centralized), to a divisional strategy (decentralized) or rather to line management (federated) (Webb et al., 2006).

Figure 2.2 shows the 2 different organizational models which may influence the choice between a centralized and decentralized organization.

	Centralized	Decentralized
Business strategy	Cost focus	Innovation focus
Business governance	Centralised	Decentralised
Organisation size	Small	Large
Information intensity	Low	High
Environment stability	High	Low
Business competency	Low	High

Figure 2.2 Determinants of centralized/decentralized IT organization (Van Grembergen, De Haes & Guldentops, n.d, pp. 25)

For Van Grembergen et al.(2004), it is the Board or executive management who is to communicate, in a clear manner, what the different roles and responsibilities of IT governance are and assign accountabilities for the tasks associated with it.
Another such structure is the *IT strategy and steering committees* which supervise areas such as audit, compensation or acquisition. In this context, the 2 committees should help the Board in all enterprise IT related matters with the difference between the 2 being that the strategy committee advises and provides input to strategic IT issues while the steering committee runs day-to-day operations of IT service delivery.

2.4.3 Processes

Processes help in shifting from governance areas to management ones and refer to the actual implementation, monitoring and control of policies and guidelines established at corporate level. This is accomplished by using methods, frameworks and procedures specialized in translating high-level IT governance objectives in detailed agreements, measures , methods, procedures and indicators.
Among the processes mentioned by Van Grembergen et al.(2004, pp.25), *balanced scorecards* link a firm's financial evaluation to measures concerning customer satisfaction, internal processes and innovation. In the context of IT, the authors have developed the IT balanced scorecard which they describe as :

> building the foundation for delivery and continuous learning and growth (future orientation perspective) is an enabler for carrying out the roles of the IT divisions' mission (operational excellence perspective) that is in turn an enabler for measuring

up to business expectations (customer expectations perspective) that eventually must lead to ensuring effective IT governance (corporate contribution perspective)

Van Grembergen et al.(2004) continue with mentioning *strategic information systems planning* which is concerned with business-IT alignment, positioning of IT as an enterprise advantage, management of IT resources, technology policies and architectures. They also refer to *service level agreements* which define the accepted levels of service by users and the key performance indicators defined to measure them. Regarding the existence of processes for IT investment decisions, *information economics* refers to a scoring technique based on the return on investment of a project and other "*non-tangibles*" (Van Grembergen et al., 2004, pp.28) which are considered as useful in the evaluation and selection of IT projects.

2.4.3.1 COBIT

Control Objectives for information & related activities (COBIT) comprises the resources needed for adopting an IT governance framework (Afzali, Azmayandesh, Nassiri & Shabgahi, 2010, pp.47), the purpose of the framework being to "*provide management and business process owners with an information technology governance model that helps in delivering value from IT and understanding and managing risks associated with it*". COBIT focuses on providing the necessary resources an organization needs to accomplish its business functions via 4 different actions : planning and organizing for the IT processes and resources, acquiring and implementing the capabilities needed to support business programs and day-to-day operations, delivering and supporting technological capabilities, monitoring and evaluating the effectiveness of the IT service in providing value to the business (Afzali et al., 2010).

2.4.3.2 ITIL

IT Infrastructure Library (ITIL) focuses on IT service management from 2 viewpoints : organizational (people) and technical (system) : IT provides the guidelines on how to define, design, implement and maintain management processes for IT services. ITIL proposes 5 different approaches for IT management with the goal of aligning IT services to business needs and services : *Service Strategy, Service Design, Service Transition, Service Operation and Continual Service Improvement*.

2.4.3.3 The strategy alignment model

It is common in literature to mention COBIT in matters of IT governance or ITIL in terms of IT management as how well they align with the business objectives and the

strategy and how well they position themselves in the *Strategy Alignment* model (originally developed by Henserson & Venkatraman, 1993). Esmaili, Gardesh and Shadrokn Sikari (2010) mention the strategic alignment mode l (SAM) as the base in IT strategy research with a multitude of such models proposed in literature. The SAM is also one of the structures mentioned by Van Grembergen et al.(2004) in figure 2.3. The model was developed with the purpose of describing the relationship between business strategy and IT strategy on 2 axes of analysis.

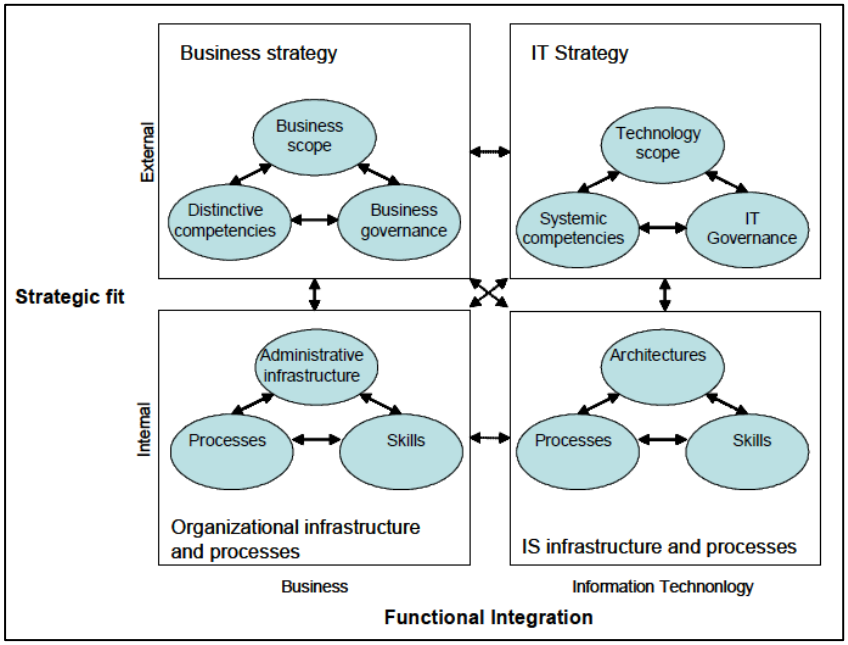

Figure 2.3 Strategic alignment model (Van Grembergen et al, n.d, pp. 8)

The 2 axes of analysis are *strategic fit* and *functional integration*. Strategic fit refers to positioning IT in an external and an internal environment. Externally, in the marketplace, IT consists of 3 domains : *IT scope, systemic competencies* and *IT governance*. Internally, in the enterprise, IT is organized from an *architecture*, *processes* and *skills* point of view. The business counterpart of the diagram suggests that business strategy should be organized following the same axes to ensure for both consistency among roles and functions as well as for synergies between the 2 counterparts.

The functional integration dimension refers to how choices are being enforced in business and IT domains accordingly. Strategic integration allows for homogenous positioning of business and IT. Operational integration refers to the coherence between constraints and anticipations of the business and the actual ability of IT to deliver.

These domains are organized along the lines of 4 positions : *strategy, technology, competitive potential and service level*. The challenge is to continuously use alignment when making decisions in any of these domains.

2.4.4 Relational mechanisms

Relational mechanisms refer to how the bridge between structures and processes is gapped in terms of interactions, collaboration and knowledge sharing. Van Grembergen et al. (2004) point to stakeholder participation and business-IT partnership as being facilitated by mechanisms such as strategic dialogue and shared learning. They also distinguish between stakeholder collaborations and partnerships on the one hand and cross-functional interactions between IT and the business on the other hand.

2.4.5 IT Governance framework summary

Figure 2.4 shows the framework of Van Grembergen et al.(n.d.) distributed accordingly in the identified elements and practices.

As a summary, *structures* are important because they show who does what, in relation to whom and the different collaborations between functions as well as analyzing what kind of investment budgets are available (investment budget, continuity budget, maintenance budget), who is responsible for each of them and which department or business unit they affect (De Haes & Van Grembergen, 2005). *Processes* refer to the effective management and implementation of IT governance structures and control frameworks (Webb et al., 2006) and to how projects are initiated, developed and maintained (De Haes & Van Grembergen, 2005).

Relational mechanisms refer to the distinction made between business and IT people as ideally, for each business role an IT role should correspond in the role charter. The importance of relational mechanisms decreases over time because while elements like training and awareness campaigns are crucial with the first implementation of the governance practices, they lose their importance as soon as the practices become repeatable processes.

Integration strategy	Structures	Processes	Relational mechanisms	
Tactics	IT Executives & accounts Committees & councils	Strategic IT decision -making Strategic IT monitoring	Stakeholder participation Business-IT partnerships	Strategic dialogue Shared learning
Mechanisms	- roles and responsibilities - IT strategy committee - IT steering committee - IT organisation structure - CIO on Board - project steering committees - e-business advisory board - e-business task force	-Balanced (IT) scorecards -Strategic Information Systems Planning - COBIT and ITIL - Service Level Agreements -Information economics - Strategic Alignment Model - Business/IT alignment models - IT Governance maturity models	-Active participation by principle stakeholders -Collaboration between principle stakeholders -Partnership rewards and incentives -Business/IT co-location	-Shared understanding of business/IT objectives -Active conflict resolution ('non-avoidance') -Cross-functional business/IT training -Cross-functional business/IT job rotation

Figure 2.4 Structures, processes and relational mechanisms for IT governance (Van Grembergen et al., n.d , pp.22)

2.5 Chapter conclusion

The IT governance framework of Peterson (2003) such as described and developed by Van Grembergen et al. (2004) is widely used and referenced across specialized literature (Lewis & Millar, 2009; Webb et al., 2006; Nassiri et al., 2009; Kuruzovich, Bassellier & Sambamurthy, 2012) and most authors build upon the structures, processes and relational mechanisms to build their own interpretations of the model with respect to different areas of research : for example, Kuruzovich et al. (2012) focus on defining and describing necessary IT governance structures.

The concept of governance in IT is, as presented, a vast concept and can refer to multiple types of elements and mechanisms, from environmental conditions to functions and scope of decision-making components. These features interact with each other by shaping and designing IT structures to enrich, establish and accomplish objectives and targets. It is complex to imagine IT governance without first establishing and identifying common corporate cross-enterprise objectives. This part however, is not explicitly mentioned in any of the existing research unless it refers to conveying the IT structures toward synergies and delivery of business value or competitive advantages.

How can these objectives be accomplished and pursued ? What are the elements interacting in the realization of these indicators which can be derived and traced back to IT needs and capabilities ? Could these elements then, in turn, be categorized to fit in one of the identified structures, processes and relational mechanisms pertaining to an IT governance framework ?

We propose to derive, for each identified general objective, smaller objectives which can be accomplished by answering a number of questions pertaining to the where, who, how, what and why, such as presented by the framework described in this chapter.

Environmental conditions in which IT frameworks can exist determine *where* we choose to place our IT function, be it low turbulence or high turbulence environments. Decision-making structures point to *who* is responsible or accountable for IT governance policymaking. *What* these policymaking themes should be and *how* they should be implemented can be exposed by looking at both the scope and fields for decision-making. Conveying these elements together builds up the foundation of IT governance as a function, which answers the last *why* question.

Once the foundation and rationale of IT governance has been set, it only remains to catalog its elements in the IT governance framework such as presented in the last section, keeping in sight how these elements should interact with each other and the

organizational environment to ensure for the accomplishment of the mission, objectives and long term strategy of the business.

3. CAPABILITY MATURITY MODELS

3.1 About this chapter

Whenever assessing an organization's standpoint with regards to its strategy, operations, investments or technology, it is important to have a starting point on which future plans and roadmaps can be build, improved and enriched to ensure for continuity and stability in an organizations overall mission and objectives. Capability maturity models build on such existing structures in order to lead the way to better steadiness and endurance in day to day operations. Building upon a capability maturity model or assessing the maturity of an enterprise means understanding how a capability model is structured, fabricated, developed and used. The most important concepts behind maturity as well as their industry definitions, principles and guidelines also ensure that the use of such a model is done in proper limits to guarantee the improvement of current performance in the passage to a superior model. The following chapter will present how such models can be erected, used and tailored to advance and rally a series of processes to the next level, while growing and progressing towards maturity.

3.2 Origin and concepts of maturity models

This chapter presents concepts of *maturity*, *performance* and *capability* as well as the origins of the first capability maturity models.

3.2.1 Mature organizations

Because Weber, Curtis and Chrissis (1994) worked on the first maturity models, they advice for first understanding the difference between a *mature* and an *immature* organization: an immature organization does not follow well-known procedures and often finds itself sacrificing aspects such as quality, reviews or testing in order to meet a schedule or remain within budget baselines. However, in spite of this focus on timely delivery, such organizations constantly find themselves going over budget and not being able to respect deadlines. By contrast, a mature organization follows a disciplined process based on value-added, clear roles and responsibilities and an infrastructure to support the process. Van Grembergen et al.(2004) mention maturity models as necessary for governance and strategy implementations because we first need to assess the current maturity level of an organization based on the identified structures, processes and relational mechanisms in order to be able to correctly design a roadmap for achieving a higher level of maturity.

3.2.2 Origins

Capability maturity models (CMM) were first developed in 1986 by the Software Engineering Institute at the Carnegie Mellon University in Pittsburgh, Pennsylvania (Paulk, 2009) and their origin stems from the inability of software developers to efficiently manage the software process. Their development was focused on encouraging a culture of software engineering by identifying critical issues to be improved based on current process maturity where each developer was to focus on a limited set of activities. The formalized concepts referring to capability maturity models were first presented in version 1.0 in 1991 by Paulk while the first official version "*Capability Maturity Model for software, version 1.1*" was released in 1993 by Paulk, Curtis, Chrissis and Weber (Paulk, 2009). Meanwhile, the software CMM has been retired in favor of the CMM Integration (CMMI) models (Paulk, 2009), which are a collection of CMM's into one framework destined for use for cross-enterprise process improvement (Chrissis, Konrad & Shrum, 2011).

One of the distinguishing features of CMM models are their *continuous* or *staged* approach. *Continuous* maturity models are based on scoring different dimensions at different levels and the summing up (or weighing) of the individual scores (Lahrmann, Marx, Winter & Wortmann, 2011). *Staged* models on the other hand require a level to comply with different processes and practices which are defined for that particular level (Lahrmann et al., 2011). Put differently, continuous approaches focus on individual process capabilities while staged approaches focus on a collection of process for a maturity level (Chrissis et al., 2011). We chose to focus on building and describing a staged approach because improving a specific process capability implies that the overall process capabilities for a maturity level has been defined beforehand. O'Regan (2011) stresses that sometimes, for a continuous approach to be successful, an organization first needs to implement a series of processes associated to a level before working on progressing a process to a different level. In order to better understand how a specific process can be improved, we need to understand what constitutes the process and how this process integrates with the other processes.

3.2.3 Capability, performance and maturity

According to Weber et al.(1994), who introduced the first capability maturity models, a fundamental concept in software development are the various differences between *process capability, process performance and process maturity*. Process capability refers to the expected results achieved by following a certain process. O'Regan (2011) reinforces the definition of Weber by stating that the fundamental notion of process refers to the tasks and/or sub-tasks necessary to accomplish a given objective. Maturity refers to the level of consistency with which processes are applied, managed and

controlled throughout different projects in the company while performance refers to the actual results achieved by following a certain process.

The initial CMM models were based on the idea that improvement comes in small, incremental steps and thus a CMM model aims at organizing these small steps into different maturity levels by defining a scale for evaluating process capability and measuring levels of maturity (Weber et al.,1994). For O'Regan (2011), a CMM provides a roadmap on how to get to a higher maturity level but it does not stipulate how processes should be done.

3.3 Capability Maturity model description

This chapter present the components of a capability maturity model.

3.3.1 Components of a CMMI

Because a CMMI is a collection of CMM models, we will use and combine elements of individual CMM models, as they were initially ascribed for software improvement processes and elements of the CMMI framework as they have been recently described by the Software Engineering Institute.

An example of how a CMMI structure is build is presented in figure 3.1 by configuring its elements in a topology.

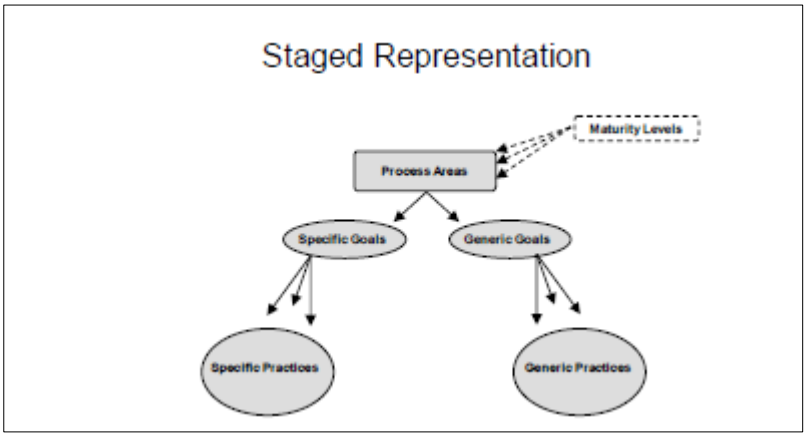

Figure 3.1 The CMMI staged representation (Team, 2010, pp.22)

Each maturity level contains *process areas* which are organized into specific and generic goals which in turn, contain *generic and specific practices* to ensure the accomplishment of these goals for each key process area. *Goals* are established for each key process area and these are used to monitor whether a key process area has been implemented

accordingly. *Process areas* indicate where an organization should focus in order to achieve process improvement (Team, 2010).

3.3.2 Process areas

Because the initial capability maturity models were focused on improving the software development process, the different process areas (called *key process areas*) for each level were specific to software processes. In figure 3.2, O'Regan (2011) provides a detailed account of each key process area ordered by maturity level. A thorough description of each key process is available in the annex A.

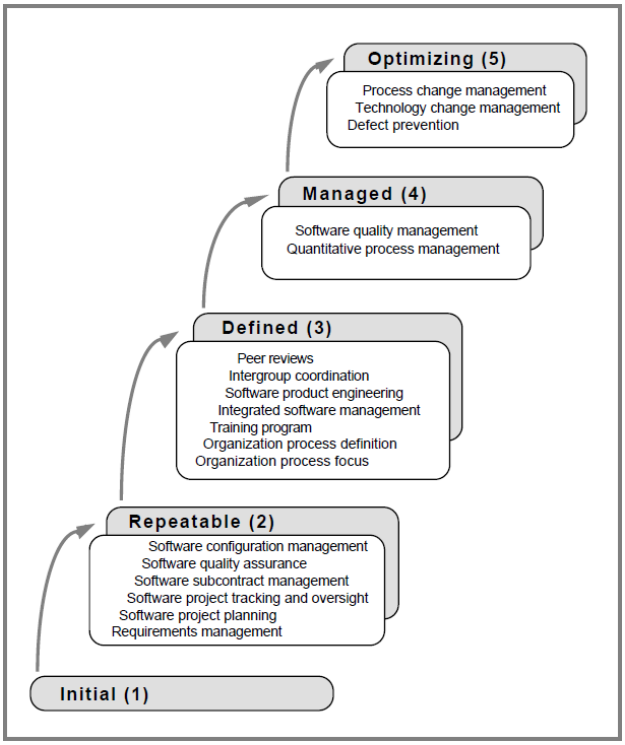

Figure 3.2 The key processes areas by maturity level (O'Regan, 2011, pp.31)

The components of a CMM are organized as follows (O'Regan, 2011): *required, expected* and *informative*. The *required components* include the generic goals (called *common features*) and the specific goals and are considered crucial to the institutionalization and implementation of the process area. The *expected components* include generic and specific practices that guide the correct and successful implementation of a process area. The *informative component* includes guidelines on how to implement these goals and practices.

3.3.3 Common features

Common features, as mentioned in previous sections, refer to how well key process areas are achieved and executed . The five common features mentioned by O'Regan (2011) are the following :

- *Commitment to perform*: institute the organizational program necessary to make a process lasting and ensure sponsorship from senior management
- *Ability to perform*: what are the resources, skills and training needed to efficiently implement key process areas
- *Activities performed*: refers to the work that needs to be done for a key process area to work properly
- *Measurement and analysis*: refers to how a successful implementation could be measured
- *Verifying implementation*: refers to potential reviews and audits as well as software quality assurance checks

3.3.4 Goals and key practices

Both goals and key practices are generic descriptions (goals) / activities (practices) which are defined according to what a key process area is expected to accomplish by its execution. Goals refer to what the key process area is expected to accomplish while key practices indicate what is needed to do in order to accomplish a specific goal without as such, indication how the goal is expected to be achieved (O'Regan , 2011).

3.3.5 Maturity levels

We wanted to mention that initially, the origin of software process improvement was associated with the work of Walter Shewhart's in the 1930's on statistical process control (O'Regan , 2011). We found Humphrey (1988) as one of the few authors what the advanced notion of statistical control is when referring to maturity models as a way for measuring process institutionalization: a process which is under statistical control will always produce the same results when it's repeated.

The SEI (Team, 2010) provides a thorough description of each maturity level in a staged approached. They describe level 1 as disorganized and unable to sustain the existence of process areas. They continue by describing this level as ad-hoc and chaotic with no formalized procedures, schedules, budgets or project plans. The crisis reaction in case of problems is to abandon all techniques and tools in place and focus on fire-fighting: this total abandonment reaction stems from the lack of experience and understanding of the consequences which come with total abandonment.

Level 2, repeatable or managed process at this level, a commitment control systems is in place and organizations have gained enough experience to be able to successfully repeat the process and results obtained so far (Team, 2010). However, because the experience is repeatable, most organizations will face challenges in their daily activities when faced with new, unprecedented experiences and projects. A process group should be set in place at this stage to focus on improving the processes in place (instead of focusing exclusively on the end product) : define the development process, identify technology needs and opportunities, review statuses and performance and report to management (O'Regan, 2011).

At level 3 (O'Regan, 2011) defined process we can say that foundations have been set and the defined process is used during crisis situations as well. There is consistency in the way projects are managed across the enterprise and these guidelines allow for tailoring and customization by project specifics. Risk management and decision analysis are implemented by following standards, procedures and criteria.

Level 4 managed (O'Regan, 2011) is characterized by the existence of a quantitative goals for evaluating key process areas and products : gathering of data over processes should be automatic and this data should be used for setting quantitative targets in order to improve measurement of productivity and quality for each process.

Level 5 optimized (O'Regan, 2011) is focused on continuous improvement, on prevention and best practices from previous projects as well as on innovation in the context of technologies and methods used.

O'Regan (2011) recommends not to skip any maturity levels as each one builds on the previous one. However, companies may astray from the standard improvement roadmap by focusing its improvements on the key process areas which are more in line with the current business goals and operations: this way, companies can benefit from actual, useful improvements. Size and current maturity levels define the time it takes to implement successive maturity models in an organization : it takes 1-2 years to implement level 2 and around 2-3 years for the following levels.

3.3.6 Domain applications for CMM's

According to O'Regan (2011) the success of the software CMM led to the development of other process maturity models such as the systems engineering capability maturity model (CMM/SE) which is concerned with maturing systems engineering practices or the people capability maturity model (P-CMM) which is concerned with improving the ability

of the software organizations to attract, develop and retain talented software engineering professionals.

Even in domains outside of systems engineering, capability maturity models are popular because they enable the organization to identify key lifecycle concepts and measurements which impact the successful implementation of business processes. Thamir and Theodoulidis (2013) mention an array of CMM models used in areas such as business intelligence, data warehousing; analytical capabilities or infrastructure optimization.

Curley (2008) has developed an IT capability maturity framework in which he identified 4 axis of management, called 'macro-processes': managing the IT budget, managing the IT capability, managing for IT business value and managing IT like a business. Curley describes each of the 4 identified key processes from a maturity level point of view : each dimension is characterized on five levels which address different perspectives of capabilities management for IT. For example, managing IT like a budget involves the existence of a sustainable economic model at level 5 while managing the IT capability pairs up with developing a technical expertise at level 3. Curley's research tested whether the level of process maturity is correlated to a value outcome. Based on the developed model, the average maturities of the 4 macro-processes turned out to be fairly good predictors of value, especially managing IT like a business proved to be the best predicator of value.

Another research in the domain of IT governance and maturity levels is AlAgha(2013) which suggests that increasing the level of IT governance maturity is best done by monitoring how IT performance is measured. He also mentions elements such as evaluation of value delivery, alignment of business and IT, monitoring of IT resources, risk and management. He continues by adding that increasing the effectiveness of IT governance is best done by appointing an IT steering committee and developing a web portal where activities related to governance are communicated as well as the existence of an IT strategy committee proved to be very helpful.

3.4 Chapter conclusion

Capability maturity models have a large applications in domains other than software development because of their methodic, efficient and organized structures which allow for a deep drill in an organizations inner working processes. Judging maturity by performance, capability and processes allows for a thorough evaluation of how well an organization is performing versus how much better an organization could be performing.

After having presented the inner workings of a capability maturity model in its original environment which is software development processes, we plan on transferring these models to other areas of an organization and building upon their original logic to construct models applicable to the problem at hand. The main objective of such a model remains to firmly ground them in an attempt to move an organization to a higher level of maturity while creating a strong, long-term competitive advantage which constitutes the basis for further improvements, advancement and progress. In the following chapters, we will show how to use capability maturity model to create strong, efficient and improved processes.

4. DATA GOVERNANCE

4.1 About this chapter

Data governance research is ambiguous in the scientific community today, mostly due to the differences existing between the concepts which form the building blocks of a governance program : data and information, governance and management, IT and business labels,… Defining and differentiating these concepts is important in understanding where a data governance structure is positioned and what the use of this term refers to. Whether these programs should be defined under IT sponsorship or loosely from such an authority is mainly determined by the contingency factors contributing to positioning an organization in both internal and external environments. For this positioning to take place, specifying a common data governance definition nevertheless proves to be crucial in determining and isolating the different elements which constitute the backbone of such programs. Identifying, defining and explaining the process layers, responsibilities and decision-making structures that come together and interact in governance topics allows for prioritizing and ranking the elements of a data governance program. These layers allow in return for tailoring to specific needs and requirements such as integration of new concepts and phenomena like big data technologies.

4.2. Concepts and theories

This chapter explains the different concepts used in data governance as well as the models and ideas used to theorize it.

4.2.1 Data and information

De Abreu Faria et al. (2013) begin their research by first differentiating between *data* and *information*. Data (pp.4437) is "*a set of symbols representing perceptions of empirical raw material*" while information (pp.4437) is "*set of symbols representing empirical knowledge, it incorporates assignment of meanings*". They point out that, in IT, these terms are used interchangeably so it is not uncommon to refer to data governance when talking about information governance. In their study, the authors opt for the latter but they explain that the choice behind using either data or information is made because the latter comprises all structured and unstructured data, as well as all kinds of data formats (video, email, documents) so one includes data governance in information governance.

Information as a concept is explained by the 3 authors by first using the resource based view (RBV) which addresses a firms competitive advantage and explains how to maintain it over time: differences in resources and capabilities between firms explain the difference in performance as not all are valued and used proportionately across the same industry; in this sense, information is considered to be such a resource. From the dynamic capabilities perspective, a competitive advantage arises not only from possession of a key resource but from correctly exploiting that resource.

4.2.2 Information and IT governance

Information governance has been acknowledged as a new concept by Van Grembergen and De Haes (2009) in the Mae's 3X3 matrix model of alignment between business and IT (Maes, 1999): more often than not, most information and communication processes are not IT dependent. Information governance in this sense, addresses the increasing importance of transforming data into information regardless of the IT-related aspects of it. Donaldson and Walker first introduced information governance at the National Health Society (De Abreu Faria et al., 2013) in 2014 for security and confidentiality arrangements in electronic information services.

Weber et al. (2009) position information or data governance as part of IT governance or comprising a part of it, while Hagmann (2013) distinguishes between the two: information governance is (pp.8) "*concerned with the way information is created, used and disposed of in order to add value to a business*" while IT governance (pp.8) "*ensures risk and compliance with IT architecture, systems and infrastructure*". Van Grembergen and De Haes (2009) also consider information governance as different from IT governance where there is a major bias on technology aspects.

4.2.3 The contingency theory

Despite the difficulty of positioning a data governance program inside or outside of IT governance ones, different authors have used the same contingency theory used in IT governance design (Otto, 2011, Weber et al., 2009) to design data governance strategies by considering internal and external specific enterprise parameters. The contingency approach is fit for use in the context of data governance because it respects the fact that each company requires a specific data governance configuration that emulates on a set of context factors. Contingencies determine which configuration is best fit for a company: by following and respecting the business goals of a company, one makes sure that data governance is not just an end in itself but that it contributes accordingly (Weber et al., 2009).

When talking about governance models, Weber et al. (2009) advises to take into account the fact that there is no data governance model that fits all companies alike and each

factor of the model should be adapted to the characteristics and specificities of an organization. This is known as the contingency theory (Weber et al., 2009) : this theory states that contingencies (e.g: size, structure,…) determine the relationship between some characteristic of the organization and its effectiveness. Figure 4.1 presents the contingency model as a variation model where contingencies are considered to be co-variation effects.

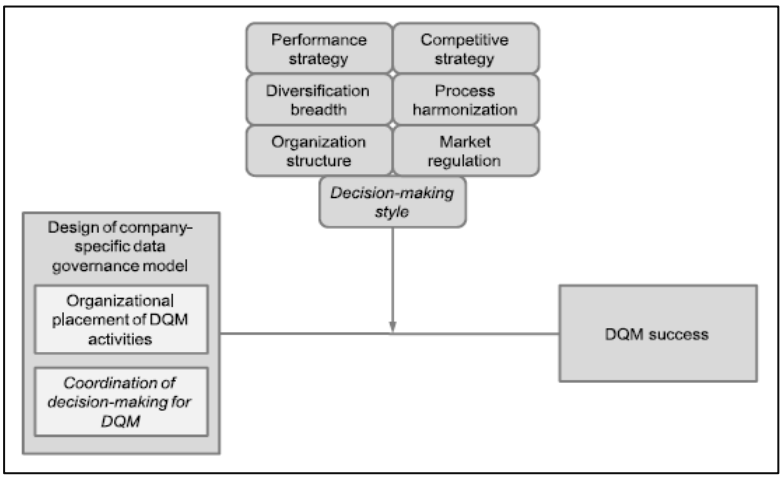

Figure 4.1 Contingencies in data governance programs (Weber et al., 2009, pp 4:16)

In the top part of the figure, we find the 7 contingency factors which contribute to the success of, in this context, a data quality management (concept which will be explained later in this chapter) program when designing organizational decision-making structures. We can notice that these contingency factors are quite diverse and differentiate between internal and external factors to a data governance model. Factors such as performance, processes and decision-making style refer to intrinsic characteristics of an organization while market regulation and competitive strategy point out to extraneous elements which could influence the way a company goes about modeling and designing for governance structures.

4.2.4 Governance and management

In literature most authors make a distinction between data governance and data management but it is not uncommon to use the 2 terms interchangeably.

Weber et al. (2009) point out to the distinction made by ISO/IEC in 2008 between the two as governance being the domain which answers the *who* and *what* questions regarding data management decision-areas while data management establishes *how* these decision will actually be implemented in practice.

Another distinction between the two is made by Khatri and Brown (2010, pp.148) : governance "*refers to what decisions must be made to ensure effective management and use of IT […] and who makes the decisions […]*" while management "*involves making*

and implementing decisions". Ladley (2012) states that managers ensure the procedures and policies are followed and adhered to while governance identifies these controls, policies, procedures, rules and guidelines.

Aiken, Allen, Parker and Mattia (2007) point out that data management has only been recognized as a discipline in the 1970's and as such, it helps transforming organizational information needs in specific data requirements. However, in his paper he includes areas such as data program coordination (which includes vision, goals, policies, and metrics) or data stewardship as data management processes which comes in contradiction with what data governance should encompass definition-wise.

We have thus chosen to include some of the processes mentioned by Aiken et al. (2007) in a data governance model as processes for which a data governance program should specify the decision-making rights and responsibilities and also because a data management program cannot exist in theory without proper governance structures. We agree that while governance specifies who will be in charge of a data management program for example, it also specifies what the elements of such a program should be. For this reason, we have included all references to data management programs in the building of our data governance model.

To support the rationale behind our choice, figure 4.2 illustrates concepts such as data governance, data management and data quality and their relations to each other for a better understanding of the differences between data governance and data management.

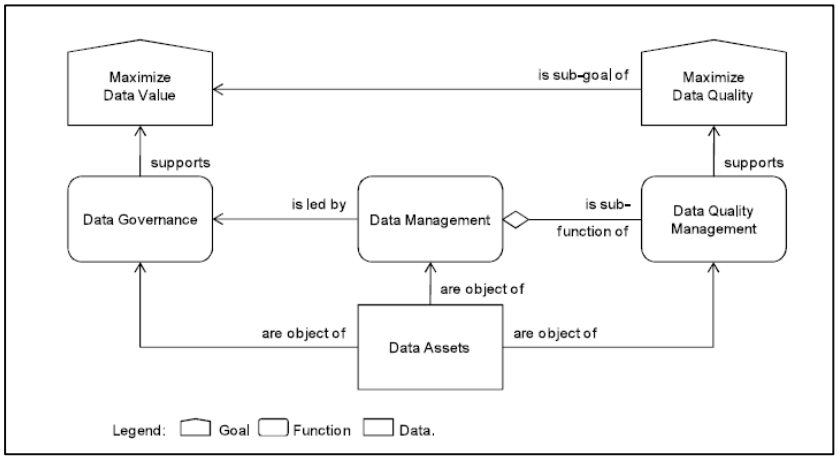

Figure 4.2 Differences in data governance and data management (Otto, 2013, pp.242)

From it we infer that data assets are addressed on 3 layers : governance, management and quality. Correctly steering and managing these data assets needs to connect each of them with a business objective, in this case, a goal : maximizing data value and maximizing data quality. Maximizing data quality is supported by data quality management (DQM) which according to Weber et al.(2009, pp. 4:2), "*focuses on the*

planning, provisioning, organization, usage and disposal of high-quality data". As such, DQM is one part of a data management program which in turn, is led by a proper data governance policy which aims at maximizing the value of data as an asset across the enterprise.

For the rest of this paper, we will use the distinction made between data governance and data management. However, we have included some processes deemed as data management processes in a data governance program because we distinguish between designing policies for these processes and actually implementing them.

4.3 Defining data governance

This chapter presents the methodology we used researching for a common definition as well as the findings associated with it.

4.3.1 Methodology of research part 1

Weber et al.(2009) notes that there is no standard definition nor in research nor practitioner community when it comes to data governance. There exist however, some definitions which are commonly shared across the scientific community and to this matter, we wanted to investigate which of these definitions comes the closest to a generally accepted (in this case, used) one cross literature.

To check for this, we conducted a literature review on data governance article, reports or conference proceedings, taking into account some borrowed elements from articles on IT governance or information governance as well as articles on data or information management. The databases we used were in particular IEEE Xplore Digital Library, the ACM digital library, Elsevier Science Direct and EBSCO host online research databases. The specific terms we searched for have varied with the findings of the research. In a first step, we searched for terms such as "*data governance*", "*information governance*", "*data management*" and "*information management*". Further, based on the findings which pointed to data governance as being part of IT governance programs, we continued the search for terms such as "*IT governance*" and "*IT data governance*" as well. Further, as we identified more and more elements pertaining to data governance programs, we expanded the search to include "*data quality management*" or "*total data quality management*" as well. We focused our research on both scientific and practitioner articles, the richness of the databases used allowed us to pick the kind of articles we wanted. While the research was focused on : 1) finding a common definition and, 2) deriving components of data governance programs, we will tackle only the first point (for the time being) in the next section.

4.3.2 Data Governance Definitions

Table 4.1 presents the compiled definitions coming from both scientific and practitioner communities divided by Author, Definition and Focus of data governance definition. The Author references the source(s) while the focus of definition is a constructed field which points out to the central point expressed in the definition.

Author	Definition	Focus of definition
Mohanty, Jagadeesh, Srivatsa (2013)	Foundational components and appropriate policies to deliver the right data at the right place at the right time to the right users	Policies, components
Tallon (2013)	Organizational policies or procedures that describe how data should be managed through its useful economic lifecycle	Policies, procedures
Otto (2011)	Refers to the allocation of decision-making rights & responsibilities regarding the use of data in enterprise	Decision-making rights, responsibilities
Weber, Otto, Osterle (2009)	Specifies the decision rights & accountabilities to encourage desirable behavior in the use of data	Decision rights, accountabilities
Khatri & Brown (2010)	Refers to what decisions must be made to ensure effective management & use (…)(decision domains) and who makes the decision (locus of accountability for decision-making)	Decision domains, accountability
Waddington (2008)	Data governance is the process of establishing and maintaining cooperation between lines of business to establish standards for how common business data and metrics will be defined, propagated, owned and enforced throughout the organization	Process, cooperation, standards, metrics
McGilvray (2007)	A process and a structure for formally managing information as a resource	Process, structure
Griffin (2005)	The process by which you manage the quality, consistency, usability, security and availability of your organization's data	Process
Fernandes, O'Connor (2009)	The high-level, corporate, or enterprise policies or strategies that define the purpose for collecting data, ownership of data, and intended use of data	Policies, strategies
Griffin (2008)	The ability to use IT to standardize data policies across the enterprise so you can gain a reliable view of the data and make better decisions	Policies

Soares (2011)	The formulation of policy to optimize; secure, and leverage information as an enterprise asset by aligning the objectives of multiple functions	Policies, objectives
Kooper, Maes, Lindgreen (2009)	Involves establishing an environment of opportunities, rules and decision-making rights for the valuation, creation, collection, analysis, distribution, storage, use and control of information	Rules, opportunities, decision-making rights
Sucha (2014)	The organization & implementation of accountabilities for managing data. Data governance includes the roles for managing data as well as the plans, policies, and procedures that control-in essence govern-data	Accountabilities, procedures
Alves Bahjat, Senra Michel, Gronovicz, Rodrigues (2013)	Information governance is a program that aims to orchestrate people, processes and technology so as to identify roles & responsibilities regarding a company's critical data inventory and, at the same time, to confer the required quality	Program, responsibilities

Table 4.1 Definitions of data governance

After having analyzed each definition and based on the focus of approach on data governance of each definition, we aggregated each element by frequency of occurrence and noted that most definitions come down to 4 elements, distributed more or less equally (we chose to exclude elements which were mentioned only once or twice and group together similar elements like processes, components and structures into processes and accountabilities into responsibilities) : policies is the most mentioned element in a definition, followed by *decision-making rights*, *responsibilities* and *processes*.

By adapting the different definitions from the table above, we have derived the following general definition for data governance : *"Data Governance encompasses the enterprise policies and processes which specify the decision-making rights and responsibilities regarding the intended use of data across the enterprise"*.

This definition is in line with what Otto (2011) defines as being the 3 data governance crucial questions one must ask before designing a data governance program :

- *What decisions need to be made regarding corporate data* ? (policies and processes)
- *Which roles are responsible* ? (responsibilities)
- *How are these roles involved in the process of decision-making* ? (decision-making rights)

We will address each of these questions separately later in this chapter.

4.4 Data governance classifications

This chapter presents the layers, practices, segments and principles associated with the practice of data governance.

4.4.1 Data quality management

Weber et al. (2009) addressed data governance from a data quality management (DQM) perspective because data governance goes hand in hand with data quality : it is not enough to have the data, this has to be high-quality in order to satisfy it's "fitness for use". Quality in this context means accuracy, completeness, consistency, relevancy and timeliness. The model they build addresses DQM on 3 layers : *strategy*, *organization* and *information systems*.

Strategy is concerned with the practical definition of a business case for data management as well as setting up a maturity assessment.

Organization is concerned with the actual implementation and monitoring of DQM initiatives. To this regards, the authors advise to take into account two design parameters : organizational structuring and coordination of decision-making. Organizational structuring is taken from IT governance research and refers to whether the IT governance design is centralized or decentralized. The centralized one places final authority to one central IT department while in the decentralized one this authority is distributed across individual business units. The coordination of decision-making structures as a second design parameter proposes two elements :

- Hierarchical models are characterized by a top-down approach where tasks are merely delegated and not discussed;
- Cooperative models on the other hand imply working in groups and making collective decisions through formal and informal coordination mechanisms.

Organizational factors are also mentioned in Tallon (2013) as one of the enablers or inhibitors in determining whether data governance is a success or failure.

The information systems layer addresses the development (logical) of a corporate data model along with the architectural design of this model and defining system support.

4.4.2 Structures, operations and relations

Other authors (Tallon, 2013) distinguish between 3 governance practices:

- Structural practices refer to IT and non-IT decision-making regarding data ownership, value analysis and cost management;
- Operational practices regard the actual execution of the data governance policy and they imply activities such as : enforcing retention/archiving policies, setting up

backup and recovery practices, access rights management, risk monitoring, storage provisioning;

- Relational practices refer to the formal/informal information flow throughout the business line regarding knowledge sharing, training, education,…

4.4.3 Outcomes, enablers, core and support disciplines

From the practitioner community, IBM (2014) proposes 4 different governance segments: outcomes, enablers, core disciplines and support disciplines. Outcomes explain and present where we want to go and what we want to achieve with data governance. Enablers refer to the organizational structures and design in place to support policies and stewardship for the governance program. Core disciplines refer to issues such as quality, security or lifecycle management. We will go over these elements later when building a data governance model. Supporting disciplines refer to classifications and data auditing activities.

4.4.4 Principles of data governance

Griffin (2010b) identifies a number of principles to be taken into consideration when developing data governance strategies : clear ownership for governance initiatives like a data governance committee or council which should decide and design data policies, procedures and standards, value recognition of data as an asset in the enterprise all the way to the C-suite level; effective data policies and procedures which should be cross functional and cross departmental; data quality and trust for the sources of data. Cheong and Chang (2007) also identified a number of critical success factors when making a case for data governance. These success factors address issues such as standards, managerial blind-spot (meaning that a program should be made fit for purpose by aligning it with the corporate strategy), cross divisional issues, partnerships or compliance monitoring.

The identified principles or success factors in literature are not homogenous and mostly point to the elements a data governance program should encompass rather than control objectives or activities to be conducted when designing such a program. These principles, while far from being generic and applicable to all forms a data governance program may come in, they can be applied on a case-by-case basis as optional practices.

4.5 Data governance processes

This chapter presents the key process areas of data governance programs and a theorized version of a data governance model.

4.5.1 Methodology of research part 2

Coming back to the distinction made by Otto (2011), this section will focus on answering the first question identified by the author, namely : *what decisions need to be made regarding corporate data* ? More specifically, we researched and identified the areas of decision in data governance programs.

Using the same methodology described in section 4.3.1 we based our research on the same literature review as previous, for both scientific sources (quite scarce regarding data governance) and practitioner sources (quite a few but less structured). We have then assembled all the different processes mentioned in these sources and based on how frequently one element is mentioned by different authors, a list of processes has been ranked by importance. It is common that the same process is mentioned more than once or in a slightly different denomination. In this case, we have chosen only one denomination for the final model. It is also the case that some processes are similar or have similar applications. In this case, the elements have been grouped together to form one process. If the elements in a group were not homogenous enough to form one process (they referred to different facets of the same general process) then they were considered as sub-processes and categorized as such. The list, along with the corresponding references is included in the Appendix B and Appendix C.

4.5.2 Data governance key processes

The elements we have identified as being the most frequently mentioned in data governance program design or initiatives are centralized in table 4.2 (references on how these processes have been aggregated and transformed into homogenous categories are available in the appendices).

Process	Sub-process
Roles, structures & policies	• Culture and awareness • People • Policies and standards • Business model • Processes & practices • Data stewardship
Data management	• Document and content management • Retention and archiving management • Data traceability • Data taxonomy • Data migration • Third party data extract • Data storage
Data quality management	• Quality methodologies and tools definition • Quality dimensions • Quality communication strategies
Metadata management	• Definitions of business metadata • Metadata repository
Master data management	• Reference data management • Data modeling • Enterprise data model

	• Data stores • Data warehousing • Data integration
Data architecture	• Data entity/data component catalog • Data entity/business function matrix • Application/data matrix • Data architecture definition
Technology	• Infrastructure • Analytics • Business applications
Security & privacy	• Data access rights • Data risk management • Data compliance
Metrics development and monitoring	• Benefits management & monitoring • Value creation quantification

Table 4.2 Data governance key process areas (details in appendix B & C)

A definition of each element is imposed for a better understanding of the identified model.

Roles, structures and policies provide, as Chapple (2013) said, the foundation for data governance programs. Roles, according to Griffin (2010b, pp.29) refer to "*ownership for governance initiatives"* while structures refer to the existence of *"fiduciary responsibility"* (IBM, 2007, pp.10) between business and IT regarding how data is governed across different enterprise levels.

Data management has many definitions associated to it and these definitions span from reference master data management, metadata management or data quality management. However, we have identified these processes as separate ones. The difference we make between these different concepts is in line with the "*Generally accepted recordkeeping principles*" (ARMA, 2015) and the concept of *records management*. We define data management practices as pertaining to (ARMA, 2015, pp.2) : "*any recorded information, regardless of medium or characteristics, made or received and retained by an organization in pursuance of legal obligations or in the transaction of business*".

Data quality management as defined by Mosley (2008, pp.11), refers to :"*planning, implementation and control activities that apply quality management techniques to measure, assess, improve and ensure the fitness of data for use*".

Metadata management is defined by Mohanty, Jagadeesh and Srivatsa (2013) as the ensemble of practices providing a homogenous definition of the data elements across an enterprise.

Master data management is defined by the DAMA Book (Mosley, 2008, pp.11) as "*planning, implementation and control activities to ensure consistency of contextual data values with a "golden version" of these data values*".

IBM (2007) refers to data architecture as the design of systems and applications which facilitate data availability and distribution across the enterprise. In order to enrich the data architecture components with sub-elements corresponding to its implementation,

we have supplemented this component with The Open Group Architecture Framework (TOGAF)-specific data architecture catalogs. The notion of a catalogue as described in TOGAF refers to an organization's data inventory which captures all data related model entities (TOGAF, 2015). Correspondingly, the concept of an data architecture definition encompasses elements like : business data model, logical data model, data management process model, data entity/business function matrix, data interoperability requirements (e.g.: XML schema, security policies).

Technology, according to Griffin (2010a) refers to the actual software and hardware components that enable the execution of data governance processes across the enterprise.

According to Tekiner and Keane (2013), security refers to protecting the information the enterprise gathers during its operations while privacy refers to clearly defining the boundaries of usage for this information.

Metrics are defined by Cheong and Chang (2007) as defining specific (baseline) measurements against which the success of a data governance program can be quantified.

4.5.3 Responsibilities & decision-making rights

The next question addressed by Otto (2011), refers to *which roles are responsible for decision-making*. To this regard, Weber et al.(2009) try to identify main activities, roles and responsibilities as well as the assignment of roles to decision areas and main activities and propose the distinctions presented in table 4.3.

Role	Description	Organizational Assignment
Executive sponsor	Provides sponsorship, strategic direction, funding, advocacy, and oversight for DQM	Executive or senior manager, e.g., CEO, CFO, CIO
Data quality board	Defines the data governance framework for the whole enterprise and controls its implementation	Committee, chaired by chief steward, members are business unit and IT leaders as well as data stewards
Chief steward	Puts the board's decisions into practice, enforces the adoption of standards, helps establish DQ metrics and targets	Senior manager with data management background
Business data steward	Details corporate-wide DQ standards and policies for his/her area of responsibility from a business perspective	Professional from business unit or functional department
Technical data steward	Provides standardized data element definitions and formats, profiles and explains source system details and data flows between systems.	Professional from IT department

Table 4.3 Set of data quality roles (Weber et al.,2009, pp.4:11)

Krishnan (2013) proposes a similar role structure in figure 4.4, composed as an organogram describing the flow of accountabilities and roles starting from an Executive Governance board which distinguishes between 2 councils : *program governance* and *data governance*. We notice the distinction made by the author between data and IT: program governance addresses IT challenges while the data governance councils focuses on data in the context of its business use.

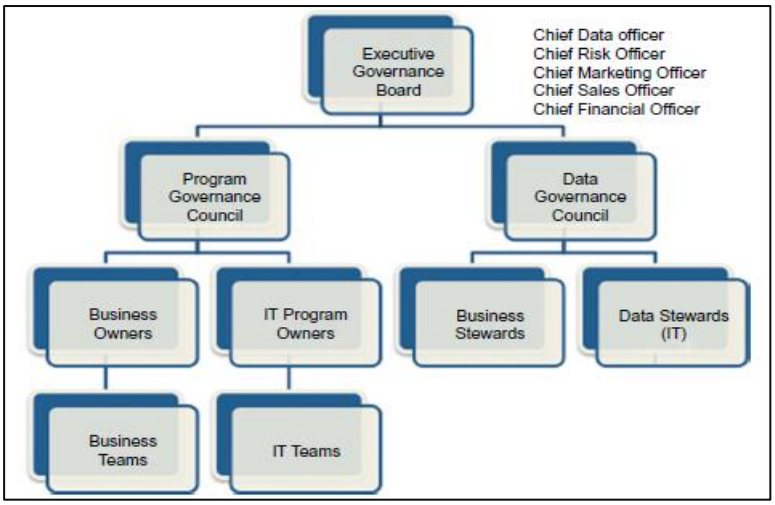

Figure 4.4 Data governance teams (Krishnan, 2013, pp.244)

This distinction is made because, as the author says, information governance is concerned with setting up overall strategies and models for data across the enterprise while program governance is concerned with implementing these strategies. The model proposed by the author is quite similar to the model proposed by Otto (2011) only more developed from a data and IT distinction point of view. The model proposed by Otto (2011) is quite simplistic and refers exclusively to a data governance program without regarding it as part of an IT department compared to Krishnan (2013) which regards data governance in a broader, corporate context. We, however, will chose the model developed by Otto (2011) for our analysis regarding data governance responsibilities and decision-making rights because of its focus exclusively on data governance programs without including IT-related issues. This will allow to better pinpoint specific roles for specific processes when designing data governance programs.

4.5.4 RACI matrix

Khatri and Brown (2010) propose a multitude of roles based on the decision domains for data governance programs : these roles span from data custodian/owner/consumer to enterprise architect or information chain manager. Some roles are very specific to a decision domain, for example, data quality demands data quality managers, analysts or

subject matter experts. However, there exists a structure in the presentation the authors make of the different roles; this structure follows a hierarchy beginning from an Enterprise Data Council **->** Data quality managers -> Data architects -> Data owners and security officers ->Data lifecycle managers.
Weber et al. (2009) assign roles to decision areas via a RACI matrix such as the one in figure 4.5 (taken from Cobit: Isaca, 2012) where each interaction is defined as Responsible (R), Accountable(A), Consulted (C) and/or Informed (I).

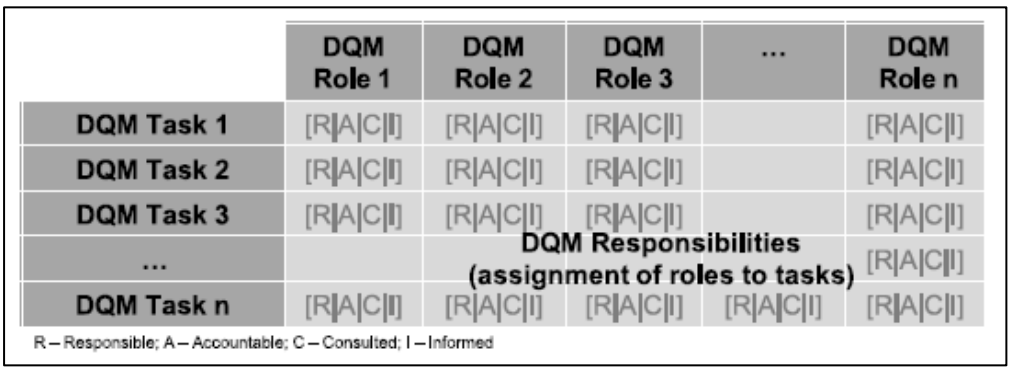

Figure 4.5 Schematic representation of a data governance model (Weber et al.,2009, pp. 4:10)

The principle behind using a RACI matrix is that each interaction fills in the cells of the matrix to depict how each role contributes to a specific process (in this case, a DQM-related task) : more than one person can be responsible for implementing a decision, however, there is only one ultimate accountable for authorizing work on a process.
If we apply the RACI matrix to the roles structure and accompanying description as presented by Otto (2011), we obtain a model such as the one presented in table 4.4.

Data governance processes	Executive sponsor	Data quality board	Chief Data Steward	Business Data Steward	Technical Data Steward
Roles, structures & policies	A	R	R	R	R
Data management	I	A	R	R	C
Data quality management	I	A	R	R	C
Metadata management	I	A	R	C	R
Master data management	I	A	R	C	R
Data architecture	I	A	R	R	C
Technology	I	A	R	C	R
Security & privacy	I	A	R	R	R
Metrics development and monitoring	A	R	R	I	I

Table 4.4 RACI matrix for our data governance model

In this sense, the data quality board should not be confounded with data quality management activities. As described by Otto (2011), the data quality board is responsible for the data governance framework as a whole across the enterprise. It is not then surprising to see that in the responsibilities assignment, the board is accountable for all decisions regarding the management and use of data across the enterprise while the Chief Data Steward, in its role of supervising both business and technical data stewards, is responsible for implementation of data management strategies and processes across the enterprise. Based on the distinction between business and technical data stewards, activities regarding data modelling or quality, metadata as well as technology and security related issues are more likely to fall under the responsibility of the technical data steward while issues regarding general data management or stewardship are implemented by the business data steward.

4.6 Chapter conclusions

We have showed that data governance strategies are mainly designed using a handful of concepts and theories which help in shaping governance related areas and processes. Presenting these concepts is useful in understanding what data governance actually "sells": data is considered as an essential asset in an organization and governance takes care that this asset is maximized, valued and used as such. This simplistic definition only resumes, of course, as we have seen, decisions on a large palette of potential policies, practices and decision-making structures.

While the data governance model we have showed in this chapter encompasses all potential elements and components such a program should cover, it is important, as with all enterprise-wide policies and practices, to focus first on elements which complement

the objectives and goals already in place in an attempt to maximize both short and long term strategies. Such a model is built step-by-step with emphasis on its most relevant components as specified by what a company/department/business unit is set to achieve. The model is complex for a complex environment but it can be tailored to smaller projects as well by prioritizing only some parts of it.

This chapter also showed the importance of having well-built and defined governance roles and responsibilities to ensure for success and industrialization of governance practices cross-enterprise. We recommend however that these responsibilities and accountabilities are defined by following the model and not the other way around. Also, each role should be mapped to each process (or a grouping of processes) in order to ensure for accountability and performance measurement.

Building upon this model, in the following chapters, we will show how new concepts and technologies can be integrated in existing governance strategies by changing the underlying assumptions and fitting them in the prevailing structures.

5. INTRODUCTION TO BIG DATA

5.1 About this chapter

Big data is indisputably one of the emerging trends regarding novel and innovative ways of utilizing data for generating more insightful decisions, increasing margins or driving operational efficiency. The complexities behind the ever growing, multifaceted data sets come in terms of new data sources and data types which need to be integrated in the actual landscape of an organization before one can "harvest" the perceived associated benefits. Defining such a new concept is challenging because big data refers to various data dimensions such as volume, variety, velocity and value. It also points to new topics and themes such as distributed processing or advanced analytics algorithms which are best explained in comparison to the current state of technology and infrastructure. Big data raises new questions as to what must we pay attention to when incorporating such new elements to our present-day systems and how these practices can be leveraged without complete disruptions in the daily usage. Thoroughly explaining what big data entails from origins and definitions to points of concern allows to build a logic understanding of its proportions before moving it into production.

5.2 Big Data : Definition and Dimensions

This chapter introduces the reader to the definitions and dimensions of big data.

5.2.1 Defining Big Data

Trying to define big data is a challenge in the academia world as a consensus on what the concept in its entirety should mean or stand for has not (yet) been reached. Zhang, Chen and Li (2013) do not categorize it as a new concept but rather as a new "*dynamic trend*". This difficulty stems for multiple reasons and most authors, while not agreeing on a definition, do agree on the multiple reasons for which a definition is momentarily lacking.

Hansmann and Niemeyer (2014) conducted a study in which they tried to both define the big data concept and characterize its dimensions based on the topics tackled by a number of articles and references on the subject. They noted that while big data has gained more and more in publication popularity (with its tipping point presumably somewhere in 2010), still no common definition of big data exists. However, they assembled a number of existing definitions from top-ranked journals and conference

proceedings and focused on whether these definitions focused more on data characteristics, IT infrastructure or methods.

We have followed the same approach in trying to reach a common definition across academia and practitioner sources. The method followed was a literature review of scientific journals and conference proceedings as well as some practitioner sources and books on the topic of big data (the same methodology and sources as the ones already mentioned in chapter 4). Consequently, we have also integrated the definitions found by Hansmann and Niemeyer (2014) but chose to drop the distinction made on definition focus because more often than not, as we will further show in this chapter, big data has mostly been defined by its V's (dimensions) which we will extensively explain later in this chapter. We have thus replaced the definition focus by a dimension focus.

The purpose of the research was to reach a common definition for big data and more specifically one that encompasses the most frequently mentioned dimensions. Table 5.1 groups all definitions of the term Big data as well as their references and secondary sources, mapped against most common dimensions mentioned in the definition or inferred from the definition.

Reference author		Definition	Source	Dimension
Hansmann Niemeyer (2014)	**&**	The exploding world of big data poses, more than ever, two challenge classes : engineering-efficiently managing data at unimaginable scale; and semantics finding and meaningfully combining information that is relevant to your concern (…) In this big data world information is unbelievably large in scale, scope, distribution, heterogeneity, and supporting technologies	Bizer et al. (2011)	Volume
Hansmann Niemeyer (2014)	**&**	(…) data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization techniques	Chen et al.(2012)	Volume, variety
Hansmann Niemeyer (2014)	**&**	"Big Data" refers to enormous amounts of unstructured data produced by high-performance applications falling in a wide and heterogeneous family of application scenarios: from scientific computing applications to social networks, from e-government applications to medical information systems, and so forth	Cuzzocrea et al.(2011)	Volume, variety
Hansmann Niemeyer (2014)	**&**	Recently much good science, whether physical, biological, or social, has been forced to confront –and has often benefited from – the "Big Data" phenomenon. Big Data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advantages in data recording and in storage technology	Diebold et al.(2003)	Volume, veracity
Hansmann Niemeyer (2014)	**&**	Data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time	Jacobs (2009)	Volume
Hansmann Niemeyer (2014)	**&**	Data that's too big, too fast, too hard for existing tools to process	Madden (2012)	Volume, velocity
Hansmann Niemeyer (2014)	**&**	Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze	Manyika et al. (2011)	Volume

Wielki (2013)	Big data it's a characterization of the never-ending accumulation of all kinds of data, most of it unstructured. It describes data sets that are growing exponentially and that are too large, too raw or too unstructured for analysis using relational databases techniques Data sets so large, so complex or that require such rapid processing (…) that they become difficult or impossible to work with using standard database management or analytical tools	Wielki (2013)	Volume, variety, velocity
Khan, Uddin, Gupta (2014)	A form of data that exceeds the processing capabilities of traditional database infrastructure or engines		Volume
Mohanty, Jagadeesh, Srivatsa (2013)	Extracting insight from an immense volume, variety & velocity of data, in context, beyond what was previously impossible	IBM	Volume, variety, velocity
Alves De Freitas, Senra Michel, Gronovicz, Rodrigues (2013)	Big data is a new term, used to describe the great volume of information that is originated from various channels, such as companies' traditional systems, the Internet and the social networks, among others, and use this information to analyze & understand people's behavior		Volume, variety, value
Buhl et al. (2013)	A multidisciplinary and evolutionary fusion of new technologies in combination with new dimensions in data storage and processing (volume & velocity), a new era of data source variety (variety) and the challenge of managing data quality adequately (veracity)		Volume, velocity, variety, veracity
Chen, Mao, Liu (2014), **Hu, Wen, Chua, Li (2014)**	Datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time A new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, analysis and discovery	Apache Hadoop definition reference IDC (2011) definition reference	Volume, Variety, Value, Velocity
Ohata, Kumar (2012)	Typically the explosion of user transactional data that reveal the patterns and behaviors of consumers		Variety
Bedi, Jindal, Gautam (2014)	The collection of large data sets that are very complex and voluminous in nature and it becomes difficult to process and analyze them using conventional database systems The tools or techniques for describing the new generation of technologies & architectures that are designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery and/or analysis		Volume, variety, value
Hu, Wen, Chua, Li (2014)	Datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze Data volume, acquisition velocity, or data representation" which "limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant horizontal scaling for efficient processing	Mckinsey (2011) NIST (2012)	Volume, velocity
Ebner, Bühnen, Urbach (2014)	Phenomenon characterized by an ongoing increase in volume, variety, velocity, and veracity of data that		Volume, variety,

	requires advanced techniques and technologies to capture, store, distribute, manage and analyze these data	velocity, Veracity

Table 5.1 Big data definitions in literature

Judging by the dimension characteristic, most big data definitions refers to volume and variety of data sets with variations regarding how fast data is produced (velocity) or the insights derived and used from processing and analyzing such data (value). The volume dimension is mentioned with a frequency of 16 times with the variety dimension mentioned in 10 definitions. Following, we have velocity mentioned in 7 definitions, value mentioned in 3 and veracity mentioned 3 times.

Before settling on a common definition on what big data is, we also favored a research of its most commonly mentioned dimensions not only based on definitions but also based on the body of research. We will discuss the dimensions aspect in the next sections before attempting to provide a definition and present our own dimensions model.

5.2.2 Dimensions model in theory

The dimensions model was first published by Gartner as a 3 V's model (Morabito,2014) : *volume*, *velocity* and *variety* but in 2011, an IDC report added the value dimension to the initial model. This latest dimension highlighted the most critical aspect of big data : discovering/mining value. A lot of definitions from the practitioner community (such as IBM mentioned in the previous table) use the original Gartner 3V model, although new dimensions such as veracity or validity are added to fit the different facets of research in big data (Bedi, Jindal & Gautam,2014).

The reasoning behind the attempt to structure an all-general dimensions models stems from the variety of dimensions which are continuously proposed both in the academia and the practitioner community. For example, Bedi et al. (2014) added to their 7V dimension model a 3C sub-dimension consisting of attributes such as *complexity*, *cost* and *consistency*. It is however important to focus and keep only the most commonly referenced dimensions of big data as a general concept. This can ease the implementation and deployment of an incipient big data project as it only steers focus on the first and foremost traits of the concept. Additionally, dimensions such as variability or validity, while important to mention and take into account when dealing with complex, sensitive information (such as financial consumer data for example), can very easily be integrated in the general dimensions like velocity (peaks in data recording are correlated to speed of data flows) or veracity (data can be valid but not necessarily truthful). Unlocking new levels in the big data journey will allow to further add or remove dimensions based on how relevant information complements the actual business needs.

5.2.3 Dimensions model research

We wanted to check whether the dimensions we have identified in the most common big data definitions correspond to the dimensions most frequently mentioned in big data literature. During the same literature review as previously mentioned, we noted the most common dimensions mentioned not only in the definitions but also in the body of the research papers as a potential dimension model. Some authors did not mention a particular dimension as part of the definition but did mention the dimensions model in their research. We have thus, as such, grouped the authors mentioning the same dimension(s) and counted which ones were the most frequently mentioned.

Table 5.2 groups the frequency count per researcher and per dimension for each of the sources we used in our literature review on dimensions.

Big data dimensions	Reference research	Frequency count
Volume	Buhl et al. (2013), Morabito (2014), Chen, Mao, Liu (2014), Katal, Wazid, Goudar (2013), Ali-ud-din Khan, Uddin ,Gupta (2014), Liu, Yang, Zhang (2013), Bedi, Jindal, Gautam (2014), Hu, Wen, Chua, Li (2014), Ebner, Bühnen, Urbach (2014), Zhang, Chen,Li (2013)	10
Velocity	Buhl et al. (2013), Morabito (2014), Chen, Mao, Liu (2014), Katal, Wazid, Goudar (2013), Ali-ud-din Khan, Uddin, Gupta (2014), Liu, Yang, Zhang (2013), Bedi, Jindal, Gautam (2014), Hu, Wen, Chua, Li (2014), Ebner, Bühnen, Urbach (2014), Zhang, Chen,Li (2013)	10
Veracity	Buhl et al. (2013), Morabito (2014), Ali-ud-din Khan, Uddin, Gupta (2014), Bedi, Jindal, Gautam (2014), Ebner, Bühnen, Urbach (2014),	5
Variety	Buhl et al. (2013), Morabito (2014), Chen, Mao, Liu (2014), Katal, Wazid, Goudar (2013), Ali-ud-din Khan, Uddin, Gupta (2014), Liu, Yang, Zhang (2013), Bedi, Jindal, Gautam (2014), Hu, Wen, Chua, Li (2014), Ebner, Bühnen, Urbach (2014), Zhang, Chen,Li (2013)	10
Accessibility	Morabito (2014)	1
Quality	Morabito (2014)	1
Value	Chen, Mao, Liu (2014), Katal, Wazid, Goudar (2013), Ali-ud-din Khan, Uddin, Gupta (2014), Liu, Yang, Zhang (2013), Bedi, Jindal, Gautam (2014), Hu, Wen, Chua, Li (2014), Zhang, Chen, Li (2013)	7
Variability	Katal, Wazid, Goudar (2013), Bedi, Jindal, Gautam (2014)	2
Complexity	Katal, Wazid, Goudar (2013)	1
Validity	Ali-ud-din Khan, Uddin, Gupta (2014)	1
Volatility	Ali-ud-din Khan, Uddin, Gupta (2014)	1
Viability	Bedi, Jindal, Gautam (2014)	1

Table 5.2 The most frequently mentioned dimensions of Big Data

When characterizing big data by its dimensions model, as is often the case in literature, the initial 3V Gartner model is the most commonly referenced. However, value appears as a runner-up for the fourth dimension, with veracity as fifth. We also notice some marginal dimensions such as volatility and quality but these dimensions are mostly linked to the subject of research of the paper: research on big data technologies and infrastructure such as Ebner et al. (2014) use a simplified 3V or 4V model because the topic deals with the theoretical aspects of the big data concept; research on innovation, opportunities and potential challenges big data brings about, such as Morabito (2014) tend to present a 360° picture of big data as a phenomenon and not as a concept, exploring thus all facets and characteristics through the adding of extra-dimensions.

5.2.4 Proposed definition and dimensions model

Based on the results presented in table 5.1 and 5.2, we have derived a potential big data dimensions model, as a 4(5)V dimensions model :

- Volume : data volumes and dataset size;
- Variety : structured, semi- and unstructured data;
- Velocity : speed of data creation;
- Value : the outcome of data processing;
- (Veracity) : truthfulness of data and how certain we can (or not be) of it;

The reason for which the model contains either 4 or 5 dimensions stems from the frequency of use for the veracity dimension : it is mentioned in 50% of the cases while the other identified dimensions are present in over 70% of the cases. As a dimension, veracity is important in assuring the data we use is the authentic data but as it relies entirely on the security infrastructure deployed (Demchenko, De Laat & Membrey, 2013), we will leave veracity as a dimension to be considered when dealing with pure big data infrastructure or technology issues. It is not a coincidence that veracity appears in 50% of the cases as most research papers currently available deal with big data as a technology and not as a solution. For this reason, we include veracity in our model to be considered only when the nature of the project to be deployed involves dealing with advanced security infrastructure issues. When defining a big data roadmap consisting of most important use cases, then we advise the use of the 4V dimension model. New dimensions can be added along a project if the need arises to treat challenges which could not be previously forecasted or to accommodate new emerging trends in the theory.

In line with our findings, the definition we agreed upon to use for the remaining of this research is a combined version of the IDC definition presented by Hu et al.(2014) and the Diebold et al.(2013) definition presented by Hansmann and Niemeyer (2014) :

A new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, analysis and discovery of potentially relevant data (veracity)

The reason of choice behind this definition is 3 folded: 1) it refers to big data largely from a technological and architectural point of view without solely focusing on data characteristics; 2) it incorporates all the identified dimensions in our research which positions it in line with our own findings and 3) it underlines the economic potential of extracting value from big data.

5.3 Big data features

5.3.1 Origin and size

The "Big Data" term was first coined by Doug Laney, an analyst at the META Group (now Gartner), in 2001, in an annual report regarding emerging technologies called "*3-D Data management: controlling data volume, velocity & variety*" (Bohlouli, Schulz, Angelis & Pahor, 2013). Gartner has meanwhile, posited big data at its tipping point nowadays, with a broad adoption to be expected in the next 5 years (Buhl et al., 2013).

Chen et al. (2014) note that in 2011, the total amount of data copied and created in the world was 1.8ZB, roughly 10 at the power of 21 bytes. This number has, at that period, been estimated to increase nine-fold in 5 years. Hu, Wen, Chua and Li (2014) reference an IDC report (2012) which predicts that from 2005 to 2020, global data will increase 300 fold from 130 exabytes to 40,000 exabytes which translates into data doubling every two years. They also expected, that by 2012, this data will be 90% unstructured (Ebner et al. 2014). Wielki (2013) quantified that in 2012, 2.5 exabytes of data were created every day, this amount being estimated to double every 4 months onwards.

5.3.2 Early trends

Understanding the starting point of big data in today's digital landscape is an important step in being able to categorize the data deluge it brings with. To this matter, Wielki (2013) has identified a number of trends which have contributed to the development of the big data phenomenon such as:

- the growth in traditional transactional databases which forced companies to collect more and more data about the customer as a potential competitive advantage but also the increasing expectations from customers regarding products and services;
- the growth in multimedia content which constitutes more than half of internet traffic data;

- the development of the Internet of Things (IoT) where devices communicate with each other and exchange information without human interference;
- social media and social networking information.

5.3.3 Data sources

As a result of these developing trends, Georges, Haas and Pentland (2014) have identified and categorized 5 key sources of big data:

- Public data as data held by governments and governmental bodies as well as national & local communities over topics such as : transportation, energy use and health care;
- Private data as data held by private businesses, NGO's and other individuals like consumer transactions, RFID tags, mobile phone usage, website browsing;
- Data exhaust as data which is passively collected like internet search logs, telephone hotlines, information-seeking behavior;
- Community data like consumer reviews, voting buttons, feeds ;
- Self-quantification data as quantified information about an individual's behavior and preferences.

Another common distinction made between categories of (big) data is a taxonomy proposed by Oracle, used both by Khan, Udding and Gupta (2014) and Liu, Yang and Zhang (2013) :

- Traditional enterprise data : CRM systems, ERP, Web stores, General ledger data;
- Machine/sensor generated data : Call detail records, Weblogs, Digital exhaust, Trading systems;
- Social data : Posts, Tweets, Blogs, Emails, Reviews.

Ebner et al.(2014) divide data in 4 different classes:

- External structured data (GPS location data, credit history,…);
- Internal structured data (CRM, ERP, inventory systems,…);
- External unstructured data (Facebook & Twitter posts,…);
- Internal unstructured data (sensor data, text documents,…).

5.3.4 Traditional data and big data

Understanding the new data sources also means understanding the difference between traditional data and big data. Hu et al. (2014, pp.654) have used such a comparative

model in figure 5.1 to distinguish big data on all 4 dimensions (volume, variety, velocity, veracity), with structured data being centralized while semi- and unstructured data being fully distributed.

	Traditional Data	Big Data
Volume	GB	constantly updated (TB or PB currently)
Generated Rate	per hour, day, ...	more rapid
Structure	structured	semi-structured or un-structured
Data Source	centralized	fully distributed
Data Integration	easy	difficult
Data Store	RDBMS	HDFS, NoSQL
Access	interactive	batch or near real-time

Figure 5.1 Comparative model between traditional data and big data (Hu et al.,2014, pp.654)

Morabito (2014) makes a similar distinction between stocks and streams: digital data streams (DDS) and big data are different because big data is more or less static and has as main use to be mined for insight. Digital data streams on the other hand evolve over time dynamically and call for immediate action. This distinction spans also in the scope and target of decision-making: DDS is more suited for marketing and operations when the impact of the reaction is high while big data can be used more for strategy, long term decisions and business innovations. Hu et al. (2014) make the same distinction only between streaming and batch processing : streaming processing relates to using data in real time in order to derive insights and results and re-insert them back into the stream while batch processing implies first storing the data and then analyzing it which makes data more static.

The different distinctions made in literature between data sources and types seem to converge to a consensus on using two axes for classification : whether the data is internal or external to a company and whether the data is structured or unstructured. We advise on using this distinction as presented by Ebner et al. (2014) because it is intuitive and simple to implement and because it encompasses and integrates the other distinctions as well : internal structured data can include stocks of data fitted for batch processing while external unstructured data can include digital data streams.

5.3.5 Themes

Big data is an extensive subject as it can refer, as has been shown in the different definitions and research, to multiple facets of one phenomenon. Bohlouli et al. (2013)

have distinguished between different factors and strategy points for big data lifecycle phases : as such big data can refer to storage and integration, as it can refer to use and technologies such as analytics and infrastructure but it can also span to management and organization issues such as investment in appropriate human resources.
Because the subject of big data is quite extensive, Hansmann and Niemeyer (2014) have chosen to research it by using topic models. They used the Webster and Watson approach and applied a structured literature review in order to validate the derived dimensions of big data and then apply topic models to enrich these dimensions accordingly. It is important to note that big data "dimensions" term as it is used here should not be mistaken with the 3V/4(5)V model as areas of interest which characterize data as such but rather as the most common research subjects on the topic of big data. For this reason, to avoid confusion, we will use the term "theme" to name and discuss these dimensions.
In their research, the authors have thus derived 4 themes on big data:

- A data theme referring to the amount and structure of data;
- An application theme referring to how the insight gained from data is applied to the business environment;
- An IT Infrastructure theme which refers to the tools and databases used to store and manipulate data;
- A methods theme referring to the analysis tools used for (big) data processing;

5.3.6 Technologies

In the same line of research, from an IT infrastructure view, Liu, Yang and Zhang (2013) have sketched big data through the use of the different technologies, either for data management and analytics or infrastructure. In their paper called "*A sketch of big data technologies*" they explain what big data is from a pure technology theory approach by highlighting points of interest when delving into technical details about big data :

- Technology-wise, big data processing in similar to traditional data processing with a difference residing in the fact that big data processing can use parallel processing such as MapReduce which first splits and then merges back the data
- From a data acquisition point of view, big data technologies use some specific collecting methods for system logs such as : Chukwa (Chukwa, 2015), Flume (Cloudera, 2014), Scribe (Facebook, 2008). These tools are based on a distributed architecture and thus can record hundreds of MB per second. Network unstructured data collection is done by using bandwidth management technologies such as DPI which can support images, audio & video

• Data preprocessing is done the same way as traditional data using Extract-Transform-Load (ETL)

• Data storage is different for big data, the logic being the use of thousands of cheap PC's in order to save and process data. There are actually 2 known file storage technologies for Big data which are Google File System (Ghemawat, Gobioff, Leung, 2003) and Hadoop Distributed File System (Borthakur, 2012). These technologies use a master-slave control node which means that it's only the host node that receives the instructions and metadata while the slave nodes takes charge of data storage

• Database management technologies are not relational (or not to the same extent) anymore but range along different structures such as column-storage technologies and NoSQL databases

• Typical data mining activities are done by using Hive (Apache Hive, 2015) and Mahout (Apache Mahout, 2015)

5.3.7 Architecture framework

Complementing the research from Hansmann and Niemeyer (2014) on the 4 topics of interest and expanding the IT infrastructure approach taken by Liu, Yang and Zhang (2013), Tekiner and Keane (2013) propose a big data framework based on 3 stages : choosing the correct data sources (stage 1), data analysis and modelling (stage 2), data organization and interpretation (stage 3).

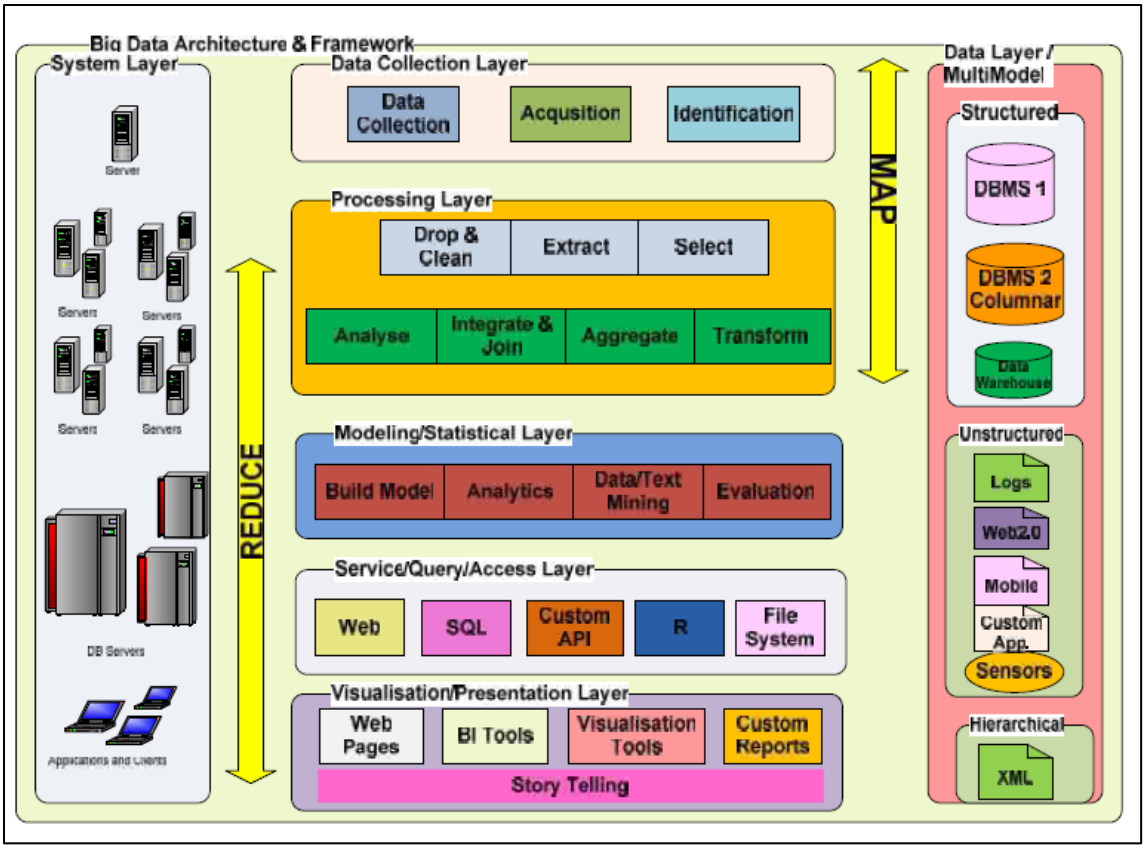

Figure 5.2 Big Data architecture & framework (Tekiner & Keane ,2013, pp. 1497)

Figure 5.2 expands these stages to constitute 7 enterprise layers which are the organized on two axes Map and Reduce. The system layer contains the platform infrastructure capable to integrate, manage and compute large data volumes. The data layer/multi-model identifies sources and types of data used in processing and analysis. The data collection layer is concerned with transforming data into information by integrating and correctly classifying it. The processing layer applies then the necessary data transformation and preparation before applying analytics and predictive models. The modelling/statistical layer turns data into intelligence by applying algorithms and calculations meant to derive useful insight. The service query/access/layer is necessary in order to map the data to the relational target model which is not directly possible for data sources available in big data applications. The visualization/presentation layer coordinates the output of the process in a clear and precise way for the business users.

5.3.8 Strategies for implementation

Ebner et al. (2014) follow a more data and methods oriented approach and propose the following 4 strategies for big data implementations:

- Relational Database Management Systems (RDBMS) are suited for approaching big data as long as data does not have to frequently be loaded into the system and exclusively for structured data. The authors reference a study by BARC where it is shown that 89% of companies rely on RDBMS when approaching Big Data compared to less than 20% who use pure big data solutions;
- MapReduce and DFS systems are fit for loading and analyzing unstructured data like text files and Facebook posts (compared to a data warehouse) but are not actually suited for an environment with frequently changing patterns and models because of the complexity of writing MapReduce code (compared to an ad-hoc query in SQL for example). These integrations and solutions are also correlated with high costs (not because of the license in itself cause open source) for migration, consulting and training efforts;
- Hybrid approach consists of integrating MapReduce capabilities for unstructured data with RDBMS engines for query optimization. However this strategy does not seem to perform better when compared with disparate strategies and it is also more expensive (usually in hardware because more processing power and storage are needed). Such examples are : HadoopDB, Oracle in-database Hadoop, Microsoft Polybase or Greenplum;
- Big data analytics as a service**:** the infrastructure for big data is hosted in the cloud which allows for economies of scale and better integration with current enterprise solutions (e.g : Cloudera). However, issues such as security and encryption are not fully

tackled with as well as integration between the cloud and the company-internal infrastructure;

5.4 Big Data projects

5.4.1 Financial valuation

Big data projects bring about challenges in the actual business landscape because they involve some specific characteristics one must take into account before and while setting up such a project. If these challenges differ in their nature and focus, there is not much doubt about what big data can actually bring about as a financial advantage.

The big data phenomenon has had some significant advances in the past few years with many companies harnessing its economic potential, estimated by McKinsey in Hu et al.(2014) at around $100 billion potential revenue for service providers from personal location data to $300 billion expected value over the next 10 years for consumer and business end-users. Ebner et al.(2014) quantified the financial value of big data strategies at $300 billion annual potential for health care, $250 billion annual potential for the public sector and e-governments and around 60% potential increase in operating margins for e-commerce, marketing and merchandising.

5.4.2 Cost, privacy and quality

Different authors have identified different challenges or characteristics of successful big data projects. Buhl et al. (2013) identified cost reduction as one of the first traits for big data technologies, combined with Moore's law of processing power. So while new technologies like in-memory analytics might be suited for handling big amounts of data efficiently and cost-effective, these must be aligned with the existing infrastructure and business processes already in place in order to effectively integrate and profit from these advancements. Another challenge identified by the same author are the country-specific privacy concerns where a significant number of customers are unwilling to accept that data about themselves might be stored for a long time. Morabito (2013) adds that, because of this increased capacity to analyze and process unstructured data to a very low level of granularity, a lot of third-parties are involved in the process and some sensitive information might get shared. Katal, Wazid and Goudar (2013) also point out that mining and gathering information about customer behavior does not only refer to the sensitive nature of such information but also to possible discriminations which for example, social media behavior can make, much of which people are not even aware of. Data quality is another crucial challenge, advises Buhl et al.(2013) because various data sources are exchanged between platforms and these platforms need to sync with each

other and offer one version of the truth across all channels. Morabito (2013) also adds the following challenges when deploying a big data solution:

- Data lifecycle management which addresses questions such as which data should be stored and which should be discarded;
- Energy management because data consumption, processing and storage consume more and more energy;
- Expendability & scalability because current systems should be designed to support future data size increase;
- Cooperation as different fields must come together to cooperate in harvesting the potential of big data;
- Analytical mechanisms which should be able to process masses of heterogeneous data within limited time;

5.4.3 Analytics

Analytics as a challenge to big data projects has been mentioned by multiple authors in literature. Katal et al. (2013) mention analytical challenges in the sense that not all data needs to be stored and analyzed but without such a proper analysis, we can never know which data is redundant and which is insightful. However, according to Johnson (2012), this endeavor is apparently constricted by insufficient understanding of how to use data for analytics insights or how to manage the risk associated to it accordingly.

For Georges et al. (2014), the trade-off between big data analytics and traditional analytics is the changing rigor of methodologies used in theoretical and empirical contributions : the use of the p-value of significance has to be revised because in the immense volume of data everything can be considered as significant. However, the authors advise on not developing too complicated models of analysis either because then we could fall into the trap of over-fitting the data. In this sense, it becomes also important to somewhat decrease the value of averages in analyses and in return, move the focus to the outliers because that is where critical innovations, trends and disruptions can be identified. The authors also advise on moving beyond correlations to causality by using theories and experiments with more variables than usually used in laboratory-designed scenarios.

Rajpurohit (2013) adds that nowadays there exists a struggle with the fact that analytics is seen as an IT solution and not as a partnership between data and decision-making structures : the logic behind models is left "*under the hood*".

Ebner et al. (2014) advise first on positioning analytics with regards to business objectives and answering the central question of how relevant big data analytics is to the business and how quickly we need the results of an analysis (urgency factor) then decide

accordingly on the most appropriate solution. As Hu et al. (2014) note, data mining activities must occur in real-time or near real-time in order to leverage for a competitive advantage but this requires different solutions and analysis systems which may not be applicable for every line of business.

5.4.4 Access, storage and processing

Katal et al. (2013) add data access and sharing of information as well as data storage and processing issues along with technical challenges as outlooks on big data. Sharing and using data to make more accurate decisions, in a more timely manner makes it so that former borders of competition and competitiveness between companies are threatened to become obsolete. However, the existing data is too big to be exploited in real-time, even if cloud solutions exist in place. This can be avoided by processing in storage place only, building indexes while collecting and storing and transporting only important results to computing. These aspects need however to be addressed in the context of fault tolerance and data quality issues.

5.4.5 Resources

Ebner et al. (2014) identify a number of other contingency factors in big data projects such as resources availabilities in terms of investment needed to start up and maintain a big data ecosystem or the abilities of the IT personnel in terms of the necessary skills and competencies. The latter is also mentioned in Katal et al. (2013) as school curricula's still focus on traditional computation systems while big data technologies are spreading without the necessary theoretical exploration. Another interesting contingency factor worth mentioning from Ebner et al. (2014) is absorptive capacity referring to how knowledge is utilized by the employees : if the people using a system do not understand its use or functioning, they will end up not using it to its full potential.

5.4.6 Use Cases

Big data use cases stem mostly from the industry, with players such as IBM (2014b) and McKinsey (2013) presenting complete strategies of use depending on the complexities, characteristics and availability of data.

IBM (2014b) identified 5 major big data use cases, organized accordingly by data dimensions, types, sources and expected goals associated with their implementation. McKinsey (2013) created a *Big Data & Advanced analytics pyramid,* organized by types of data and distinguishing between data in motion and data at rest, which is similar to Morabito's (2014) distinction between data streams and data stocks.

We have paired the previously identified sources and types of data with the most frequent use cases mentioned by IBM and McKinsey , so for example, structured types of sources such as transactional databases can be dealt with " at scale" (McKinsey, 2013) for pricing, lead generation or customer experience campaigns while unstructured ones can, for example, be transformed into structured data and integrated in campaigns aimed at cross channel data integration.

SOURCE	TYPE OF DATA	
	STRUCTURED	**UNSTRUCTURED**
Transactional Databases	• Customer experience (McKinsey, 2014) • Data warehouse modernization (IBM, 2014b) • Pricing (McKinsey, 2013) • Campaign lead generation	• Advanced marketing mix modeling identifies the impact of marketing actions on sales/churn (McKinsey, 2013) • Capturing social media buzz (McKinsey, 2013) • Shopping basket-data used to identify credit risk in the unbanked segment (McKinsey, 2013) • Advanced next-product-to-buy algorithms (McKinsey, 2013) • Cross channel data integration (McKinsey, 2013)
Multimedia content	• Data exploration (IBM, 2014b)	• Speech analytics (McKinsey, 2013)
Internet of Things	• Security intelligence (IBM, 2014b)	• Operations analysis (IBM, 2014b)
Social Media	• Advertising targeting with on-going experimentation (e.g: learning the right landing page to show to the customer) • Pricing and advertising targeting (changing price and advertising per customer)	• Enhanced 360 view (IBM, 2014b)

Table 5.3 Potential big data use cases

The matrix is only one way of integrating use cases, sources and types and it is important to note that, for example, an approach dealing with social media structured sources is not solely used for advertising targeting purposed exclusively but can also be used in data exploration analyses or cross channel data integration. The mapping between source and type of data to a use case assessment is only one example of how different sources and types can be treated and integrated from a value-added perspective to building big data strategies.

5.5 Chapter conclusions

Integrating all features of big data is a challenging mission especially because the scientific literature on successful deployments of big data projects is scarce. It is hard to pinpoint the importance of the features and challenges we have identified to actual industries and sectors, as these characteristics apply on a case-by-case basis. Giving its novelty, big data has mostly been explored as a concept or phenomenon and too little as a success story in the scientific community. Nor have any big data use cases been mentioned or developed in these researches. Use cases stem especially from the value

dimension : understanding what big data is constitutes the first step in any big data project but understanding its value added constitutes the underlying assumption which should guide every step along the way.

We have identified the most important dimensions of big data as being volume, variety, velocity and value. To this regard it comes as no surprise that our definition stems from the practitioner community as IDC (Vesset et al., 2012) has developed a number of big data industry use cases which are exclusively based on the value dimension as "smart" dimension. These use cases include pricing optimization, churn analysis, fraud detection, life sciences research or legal discovery.

In which sense is data any different from big data ? Georges et al. (2014) identify a shift in perceptions in the practitioner community from "big" to "smart": the questions is no longer how much more different big the data is compared to "small" data, but how smart the information that it provides is : the outcome might no longer be winners/losers but rather how a network interacts in order to successfully accomplish that what it wishes to accomplish. Whether the data deluge will be treated as big or smart remains to be seen by industry since in some cases, the volume dimension plays an important role in predictive activities (think about system logs) while in others, the quality of information remains crucial (think about potential fraudulent activities identified by banks).

Correctly mapping the identified big data characteristics as well as positioning big data in terms of origin and history has been the main focus of this chapter with special attention to challenges and features of big data deployments as a basis for starting big data projects.

6. DEVELOPING (BIG) DATA CAPABILITY MATURITY MODEL FOR THE BELGIAN FINANCIAL SECTOR

6.1 About this chapter

Growing volumes of data pose challenges in every sector and it is even more important to appropriately handle this data when it comes to the financial one. The financial sector reunites all characteristics previously mentioned for big data : volume, veracity and velocity. Vast amounts of records for financial transactions are generated each day but their efficient utilization remains a mystery for the appointed data owners. Given the crucial importance of banks in today's landscape, more and more regulators start to tackle the topic of proper data governance practices for their risk management practices. The unexploited "treasure" offered by the quantities of data currently owned by financial institutions sends its actors into a "gold rush" for uncovering insights and relationships never used before. This involves however, the existence of a proper environment to sustain such undertaking with the right infrastructure, architecture and policies in place to foster and develop practices which will allow for un-tapping the collected data. Regulators are already designing guidelines and frameworks to allow for accurate handling of financial records and the business structures are soon to follow if they want to keep their competitive advantages. They need to first understand how their underlying business model needs to improve in order to adapt and accommodate the ever-increasing need of data to support their core decision-making processes.

6.2 Big Data Governance

This chapter integrates the big data related concepts and technologies in the data and IT governance landscapes.

6.2.1 Big data governance models

Tallon (2013) described data governance as a reflection on how organizations see and value their data assets as well as how they plan to continue protecting these assets by investing in the appropriate technologies. In his article, he refers to big data technologies as posing a new challenge to the traditional data programs in terms of valuation, cost and governance. However, the research subject of big data governance in the scientific community is scarce: while there are some articles dealing with challenges posed by new big data technologies (Demchenko et al., 2014 being the most representative), these are loosely structured and not uniform enough to fit a close-to-standard model.

In the practitioner community, we find some sources that deal with the subject with the most prominent being the IBM Information Governance model adapted for big data (IBM, 2014). What they actually do is use the same information governance model as before (IBM, 2007) but with guiding principles regarding big data technologies. These principles refer to issues like quality or compliance which are already dealt with by most data governance programs but in the context of big data what changes is the perspective and scalability of decision domains. Information mapping and lineage become for example extremely important because the source of data will determine how valid and true the end results (of analytics, for example) will be. Scoring or using data analysis models have also changed meaning because the context dimensions of big data are no longer the same so one must first determine the tolerated level of ambiguity for example. Other such guiding principles refer to managing classifications, fostering a stewardship culture, protecting and securing sensitive information, managing classifications or increasing awareness for governance, auditing and continuous performance measurements.

6.2.2 Business and technological capabilities

Mohanty et al. (2013) identified a number of new business capabilities which are needed for big data handling such as data discovery activities for locating, cataloging and setting up access mechanisms to data sources, rapid data insight which means combining and inspecting data from multiple sources in order to spot trends and patterns more quicker or advanced analytics and data visualizations. For Tekiner and Keane (2013), the challenges of big data technologies lie mainly in data sharing decision domains because for data to be usable it first needs to be open and accessible while respecting privacy concerns and requirements which are more than ever exacerbated by the advent of geo-location or social data. They also point out to technological capabilities such as storage and retrieval which while being able to store all data, they are not able to keep short processing times with regards to the exponential growth of data unless infrastructure is being scaled up.

6.2.3 Features of big data governance programs

Demchenko et al. (2014) have identified the following features which characterize the modern changes in ICT, cloud computing and big data, with regards to governance-specific issues :

- the digitization of processes, events & products;
- automating data production, collection, storing & consumption;
- reusing initial data sets for secondary analysis;

- open access to public data and possibility of global sharing between involved groups;
- existence of infrastructure components able to support and sustain necessary cooperation and management tools;
- secure and available access control technologies to ensure a protected environment for cooperating groups.

Bahjat El-Darwiche et al. (2014) point out that any governance program should first include the formulation of a vision for the usage of data which is compatible with the public interests' approval and understanding (which data can be used, how long can it be stored, what is strictly forbidden,…) as well as fostering the knowledge and skills needed to exploit a big data environment. In terms of internal capabilities, the 3 authors mention that the main priorities for the executives should be :

- the development of an appropriate big data strategy, accentuating the value derived from pilot schemes;
- appointing an owner for big data and recruiting the right talent;
- positioning big data as an integral element in operations as well as re-orienting the culture of the organization to be more data-driven.

6.3 The financial sector

This chapter presents the challenges associated with data collection and analysis in the financial sector.

6.3.1 Financial records, information and data management

The subject of *financial records*, *information* and *data management* has received very little attention before the 2007-2009 financial crisis when it was shown how poor quality and management of financial records can lead to weaknesses in operational risk management (Lemieux, 2012). The U.S. Office of Financial Research stated that : "*Data management in most financial firms is a mess*" (Lemieux, 2012, pp.2). What the U.S. Office of Financial Research meant by its statement, continues Lemieux (2012), is that standard reference data is missing, there are no common standard designations for financial instruments and the manual correction of millions of trade records per year leads not only to inefficiencies but also to an increased risk of errors.

A report from the Financial Stability Board and the International Monetary Fund (2009) noted that :"…*the recent crisis has reaffirmed an old lesson- good data and good*

analysis are the lifeblood of effective surveillance and policy responses at both national and international levels". (Lemieux, 2012, pp.2).

In line with the policies of transparency and efficiency which banks seem to be pursuing today, qualitatively good managed records represent the foundation to monitor financial risks and counter potential threats in a timely manner (Lemieux, 2012). However, the author continues, there exists not yet a consensus on what the characteristics of such good kept financial records should be.

6.3.2 Operational and market risk

It is crucial to understand the types of risk that data collection practices can uncover and which types of risk can be properly handled by strong data collection processes. Brammertz (2012) explains the difference between market and operational risk (OR) from the optic of data-gathering processes : market risk (such as, for example, inflated prices which do not reflect the value of the underlying asset) can be mitigated by overseeing exposure while operational risk involves people and business processes at a micro, firm level.

Figure 6.1 illustrates the relationship between the 2 types of risk as an optimum between how much is done to avoid the risk and how much we are "*willing*" to incur in losses.

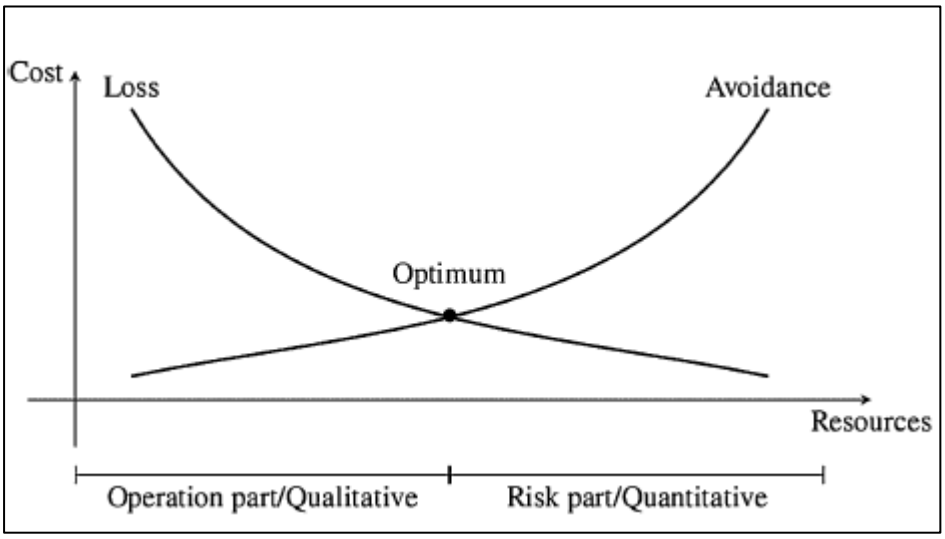

Figure 6.1 Operational and market risk (Brammertz, 2012, pp. 48)

The optimum described above differentiates between the operation part/qualitative and the risk part/quantitative. The OR quantitative, continues Brammertz (2012), is basically the market risk part, which is also the most "*uncontrollable*" one while the OR qualitative

is the controllable part, which is : "*the risk to build a huge organization, which has no chance of delivering due to principal flaws in data gathering processes*" (Brammertz, 2012, pp.49).

6.3.3 Strategic forces in financial data management

Flood, Mendelowitz and Nichols have analyzed a series of data-management related issues for the regulators and market participants as they implemented the Dodd-Frank Act (DFA). The DFA is a set of federal regulations intended for financial institutions and their customers which was adopted after the 2008 financial crisis in order to prevent a relapse of such magnitude (U.S. Congress, 2010).

The 3 authors identified 3 *strategic forces* which influence the work of data management financial supervisors : *data volumes*, *systemic monitoring* and *cognitive capacity*.

Data volumes for financial markets are growing exponentially as a coordination between back and front offices of trading firms is increasingly needed. In order to efficiently deal with increasing data volumes and types, Flood et al. (2010) recommend the transfer of evolving practices from other industries.

Systemic monitoring forces for a different angle of approach between firms and markets across the financial system and, in particular, it challenges the bilateral and multilateral contractual relationships between the network of market participants.

Building a cognitive capacity calls for a combination of "*situational awareness of the financial system*", "*decision support for policymakers*" and "*crisis response capability*".

The last 2 strategic forces look at the data management challenge from a macro-prudential scale where data validation and risk notions expand from the micro-firm level where a firm is regarded as an island shielded by unpredictable random shocks to disintermediation where the network of relationships across entities cannot be underestimated (Flood et al., 2012).

While we acknowledge the importance of looking at activities of data validation, classification, filtering or lineage across financial entities and their intermediaries, the purpose of our research is narrowed to an analysis at a micro-prudential scale and more specifically, at a firm level. We are interested in analyzing data gathering processes at an organizational level because the aggregation of these organizational levels will allow, we believe, to gather a global picture- albeit, non-systemic- of how the financial entities perform singularly in data governance programs. Aggregating information across

financial entities will then be simplified if each participant uses the same standard framework for data governance practices.

6.3.4 Data management at a micro-prudential scale

Analyzing procedures at a firm level reveals that accounting practices are still the most widely used framework for assessing a firms financial risk through its recorded assets and liabilities on the balance sheet (Flood et al., 2012). These measures, they continue, also appear in most models used for risk management practices such as : *value at risk (VaR)*,or *economic value of equity (EVE or risk-weighted assets (RWA)*). The problem with a firm-level view is that it does not take into account how aggregate imbalances across firms combine to create systemic risk ("*the volatility paradox*").

Micro-level innovations (such as, for example, the growth in derivatives market or the expansion of the trading and securitization markets), while highly regarded and encouraged, bring new types of contracts which originally are viewed as favorable and novel. However, Flood et al.(2012) show that the implications of the data management practices associated with this kind of innovations are often overlooked because of a lack of coordination between the back and front-office : innovations typically come from the front office without a solid back-office infrastructure to support them.

Scalability of data management practices is also one of the challenges for regular supervision, as, aside from growing volumes of data, data validity is of crucial importance for the financial sector : compared to the general internet data traffic, signal redundancy is not as common when it comes to financial transactions because a few corrupted digits in such a transaction could significantly alter the intrinsic value of it (Flood et al., 2012).

So far, traditional financial supervision has been firm-centric while financial information has expanded faster than the technologies needed to manage and track it (Lemieux, 2012). Managing the relationship between firms and markets across the financial sector requires systemic data collection across financial entities in a framework designed with the proper amount of governance : over-regulating bogs the system into heavy red-tape while under-regulating diminishes transparency practices (Lemieux, 2012).

6.3.5 Basel III Principles for Effective Risk Data Aggregation and Risk Reporting

Basel III is a regulatory framework released by the Basel Committee on Banking Supervision which contains rules and guidelines to reinforce and protect the global banking sector of a similar economic crisis as the one of 2007-2008 (Kindler, 2013). The framework is due for implementation in 2016. The difference with the Dodd-Frank act mentioned previously, is that Basel III requirements apply at a global level for the banking sector while the Dodd-Frank act affects only U.S. institutions (Barnard & Avery, 2011).

While Basel III has been mostly focused on calculations and computing for proper capital management by better controlling a bank's capital requirements, in January 2013, the Basel Committee introduced new guidelines regarding risk data aggregation and analysis in a document called : "*Principles for Effective Risk Data Aggregation and Risk Reporting*" (Basel Committee on Banking Supervision, 2013). The purpose of these guiding principles is to enable a quick functional access to information by accurately aggregating the information needed to respond correctly in a crisis situation (Flood et al., 2012; Kindler, 2013). The importance of such practices is highlighted by the following introductory sentence: "*One of the most significant lessons learned from the global financial crisis that began in 2007 was that banks' information technology (IT) and data architectures were inadequate to support the broad management of financial risks*"(Basel Committee on Banking Supervision, 2013, pp.8).

The main principles outlined by the Basel Committee on Banking supervision (2013) regarding practices of risk data aggregation and risk reporting are: (some principles are out of scope here, so we choose to mention only the ones which will be referenced later on):

- Governance;
- Data architecture and IT infrastructure;
- Accuracy and integrity: data should be aggregated on a largely automated basis to prevent errors;
- Completeness of data : all aggregate material risk data should be available across all possible hierarchies allowing for the timely identification of risk;
- Accessibility : access available to current and historical data;
- Adaptability : allow for on-demand, ad-hoc risk management reporting requests;
- Comprehensiveness : depth and scope of risk management reports should be coherent with the size and complexity of a bank's operations;
- Timeliness : generate and update risk reports in a timely fashion;

- Frequency : of reports per types of risk identified;
- Review : examine a banks compliance with the mentioned principles.

Some principles are out of scope here, so we chose to mention only the ones which will be referenced later on.

6.3.6 Data governance challenges in current landscape

Kindler (2013) points out that most banks nowadays adopt a *Band-Aid* approach when dealing with Basel III implementations of risk data aggregations and reporting : partly introduce solutions and applications which help consolidate a part of their data instead of integrating data across the enterprise. Skinner (2015) also points out that most banks under the $15 billion asset level do not dispose nor over the fit infrastructure nor the data management skills needed to address the principles underlined by the Basel Committee on Bank Supervision. Another problem is the externalization of the core transactional and operational processing systems which limits the access of banks to the data needed for analytical and modeling processes.

In the remainder of this paper, we plan to address the Basel III principles of risk data aggregation and risk reporting from a data-governance optic and build a twofold model : 1) a model which incorporates the Basel III guidelines in its governance practices and 2) a model capable of assessing the level of maturity of data governance processes in the financial sector.

6.4 Capability Maturity Model for Big data governance: theoretical model

This section describes the research methodology used for the development of the model as well as a description of the input used and the outputs delivered.

6.4.1 Overview of the research process

We include a visual overview of the research process which brings together components from previous chapters and binds together the different constituents which contributed to building our final model. Figure 6.2 presents the different steps of the process as well as how the different outputs and inputs interacted in the research process.

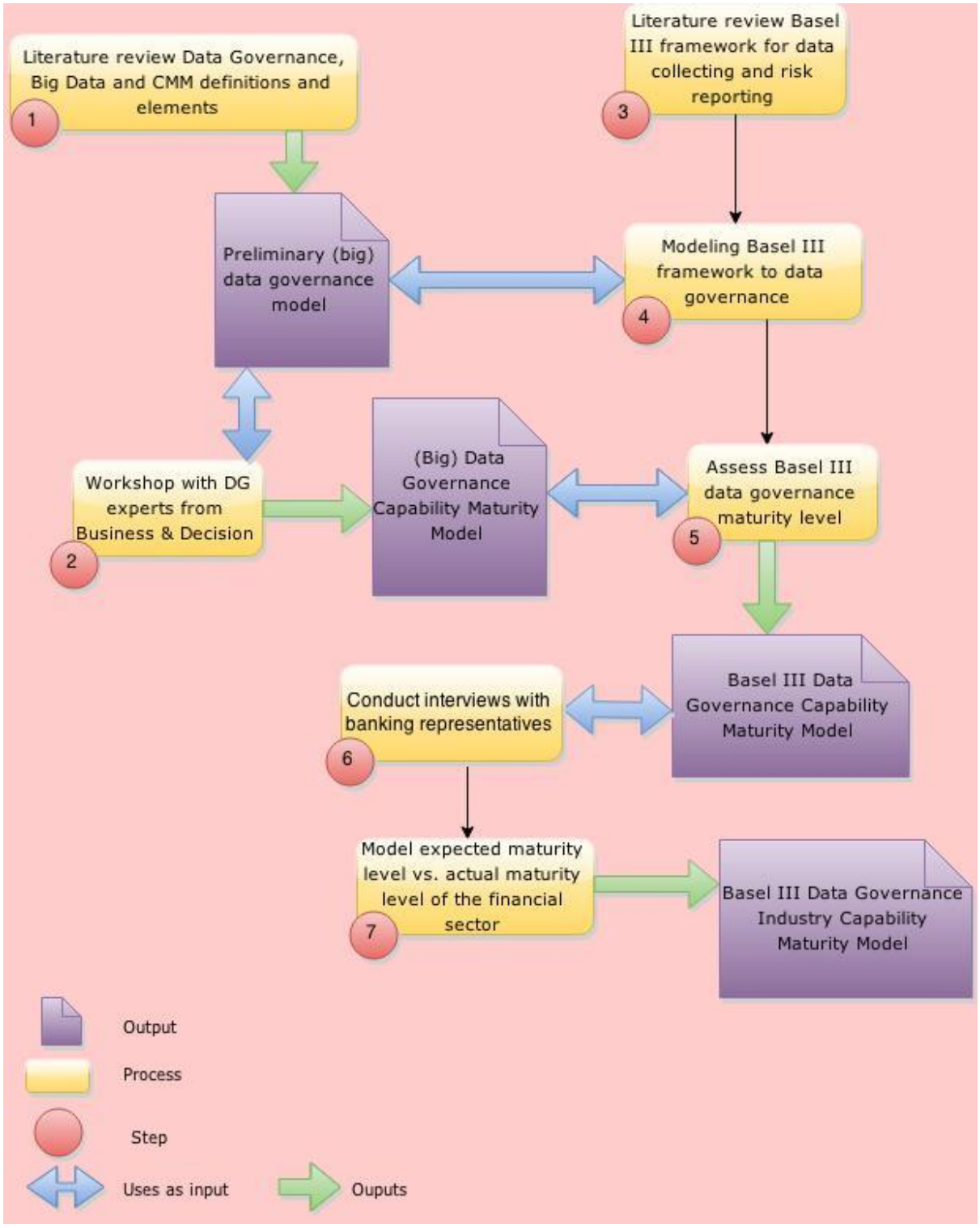

Figure 6.2 Overview of the research process

6.4.2 Research methodology part 1

In chapter 4, section 4.5.2, table 4.2 we introduced the key process areas identified during our first literature review for data governance programs and projects. The model contained 9 processes and 36 sub-processes, with corresponding definitions for each key process.

Kindler (2013) used the Basel III *Principles for effective data aggregation and risk reporting* to map the guidelines outlined by the Committee on Banking Supervision with the potential data and platform requirements. The mapping translates each guideline into the *derived technical/platform requirements* needed to implement it. Based on the mapping created by Teradata (Kindler, 2013), we then mapped the model developed in Chapter 4 to the Teradata model by translating each of the technical/platform requirements into either a key process area or a sub-processes.

The mapping has been done based on the definitions presented for each key process area in chapter 4 or by re-analyzing the definitions/references for each sub-process areas in our literature review. The mapping is then two-folded in its potential use because : 1) we test the importance of each key process area/sub-process in the Basel III guideline and 2) we help defining and better understanding what each sub-process refers to. Because of its size, we have placed the original mapping as presented by Kindler (2013) in appendix D.

6.4.3 Mapping Basel III principles to data governance key process areas

Table 6.1 presents a summarized version of the Basel III guidelines by referencing the principle and its indicated order[1], the Teradata derived requirements (also available in appendix D) and the key process areas identified in chapter 4.

Basel III Principles and Guidelines	Teradata derived requirements	Data governance key process areas/sub-processes
Principle 1: Governance, guideline 27	• Clearly defined, implemented and live data-governance policy • Clearly defined and guaranteed service levels for data processing, analysis and reporting	Roles, structures and policies (policies and standards ; processes and practices)
Principle 1: Governance, guideline 29	• Review of architectures, effectiveness, and compliance by an external and independent validation unit with specific IT, data, and reporting knowledge	Technology (infrastructure)
Principle 2: Data architecture and IT infrastructure	• Risk-architecture analysis, and reporting capabilities outlined and scaled for worst-case conditions • Infrastructure scaled to max but payment for utilization only	Data architecture Technology (infrastructure; business applications)
Principle 3: Accuracy and integrity, guideline 36 (c)	• Accuracy of reporting under stress/crisis • Automated data sourcing and aggregation, minimal manual interaction • Reconciled finance and risk data • Common data model for finance and risk • Ideally, shared data warehouse for finance and risk	Data management (Document and content management; Retention and archiving management) Master data management (data stores, data warehousing, data integration)

[1] as specified in the original document issued by the Basel Committee on Banking Supervision (2013)

Principle 3: accuracy and integrity, guideline 36 (d)	• One source of data for risk data aggregation and reporting • One source of truth	Master data management (enterprise data model, reference data management)
Principle 3: accuracy and integrity, guideline 37	• One logical data model across the risk and finance area • One business data model (access layer, etc.) across the risk and finance area	Master data management (enterprise data model, reference data management, data integration)
Principle 3: accuracy and integrity, guideline 40	• High data quality • Data-quality metrics • Automated data-quality monitoring	Data quality management
Principle 4: Completeness	• Central data warehouse with all data from all divisions within the bank • Data storing in lowest granularity level to enable aggregation across different dimensions	Master data management (data warehousing, data integration, data modeling) Data management (data taxonomy, data storage)
Principle 5 : Timeliness	• Timely import of new data to data warehouse • Rapid production of new analysis and reports (depending on criticality of results) • Intraday data on-demand import, aggregation, analysis, and reporting • System-log analysis resulting in required unstructured data-analysis tools and big data requirements	Data management (data migration, third party data extract)
Principle 6: Adaptability, guideline 48 , 49 (b)	• Flexibility in implementation of new requirements • Rapid time to market for new requirements • Ad-hoc queries • Ad-hoc analysis besides standard reporting • Forward-looking analysis and scenario calculations • Ad-hoc prediction of future risks • Interactive stress-testing across all data and risk factors of the bank across all dimensions • Drill-down capabilities to lowest granularity • Rapid visualization of ad-hoc results	Analytics
Principle 6: Adaptability, guideline 48 (c) , (d)	• Rapid business-driven change and enhancement capabilities within entire risk-aggregation value chain	Analytics
Principle 6: Adaptability, guideline 50	• Ad-hoc scenario capabilities on any set of data across the entire bank • Full business-driven flexibility in setting-up new simulations • Drill-down capabilities on ad-hoc scenario simulations	Analytics
Guideline 51	• Flexibility to send the right information at the right time to the right people	Security and privacy (data access rights)
Principle 7: Accuracy	• Reconciliation capabilities across different results	Security and privacy (data compliance)
Principle 7: Accuracy, guideline 53	• Automated reasonability and quality checks	Data quality management
Principle 8: Comprehensiveness, guideline 57	• All material risk data within the organization included in data aggregation and analysis • All transactional data produced within a bank included in the risk data warehouse	Security and privacy (data risk management, data compliance)
Principle 8: Comprehensiveness, guideline 58	• Risk results comparable across the entire organization and all divisions • One data model (logical, semantic layer, data access layer, etc.) across the organization and all divisions	Master data management (enterprise data model, data modeling) Data management (data traceability) Metadata management
Principle 8: Comprehensiveness,	• Ex-post analysis and an ex-ante simulation layer available across all risks	Analytics

guideline 60	within the bank	
Principle 10: Frequency	• Analysis and reporting frequency to match the speed with which the underlying risk may change • Capability for the reality that credit risk, market risk, and liquidity risk all depend on capital market prices and can change drastically within seconds • Intra-day risk reporting at a minimum or, better, within hours or minutes • Risk on demand in times of market turmoil	Analytics
Principle 12: Review, guideline 75	• Capability for more frequent and rapid review and testing of aggregation and analytical results by regulators • Ability to explain data and analytical results produced in the past • Rapid retrievability of historic data content and assumptions going into analysis • Temporal database design and retrieval capabilities	Data management
Principle 12: Review, guideline 76	• Review and assessment of data aggregation and analysis from external experts	Metrics development and monitoring

Table 6.1 Mapping Basel III principles to data governance processes

From table 6.1 we notice that the same elements from the model presented in Chapter 4 are mentioned more than once while mapping the Basel III principles and guidelines. Based on how frequently these key process areas/sub-processes are mentioned, we can create a ranking of the most important elements of a data governance program, applicable to the financial sector, based on the Basel III framework. The ranking frequency is presented in Appendix E.

The key process areas and sub-processes most frequently mentioned in the Basel III framework for risk reporting are the following : *analytics*, *infrastructure*, *data integration*, *data warehousing*, *data modeling*, *enterprise data model*, *data quality management*, *data management*, *data compliance*.

In the light of the findings presented so far, it is important to note that the scope of the Basel III "*Principles for effective risk data aggregation and risk reporting*" covers 4 main topics : *governance* and *infrastructure*, *risk data aggregation*, *risk reporting* and *supervisory review*, *tools* and *cooperation* (Basel Committee on Banking Supervision, 2013). The model we identified in chapter 4 is a general model which applies regardless of sector or data handling practices that a company may pursue. Otherwise put, our model is industry-free and it can be tailored to fit any sector as it is based on an extensive literature review which is grounded in theoretical research and constructed to be tailored according to practices, regulations and business models. As such, although our model does not specifically focus on risk data governance practices, its elements (key process areas or sub-processes) can be tailored the fit the scope of the Basel III framework from the perspective of the data management challenges and practices it chooses to pursue.

Among the objectives of the Basel III principles and guidelines, we have the following : "*enable fundamental improvements to the management of banks*"(Basel Committee on Banking Supervision, 2013, pp. 10). From this and from the other objectives mentioned in guideline 10, we can conclude that building a risk data governance practice as presented by Basel III does not imply starting "from scratch" but rather building, improving and enhancing the capabilities a bank already has. Kindler (2013) mentions in his whitepaper with Teradata that compliance with Basel III involves banks "*making comprehensive enhancements to their data-management processes*" (Kindler, 2013, pp.3). We will further develop and discuss this idea in the following sections.

6.4.4 Research methodology part 2

We introduced the concept of "*capability maturity models*" (CMM) in chapter 3, where we presented what a staged approach looks like, what are the process areas per maturity level and which are the common features usually used by these models.

Following the process area – maturity level description in chapter 3, section 3.3.2, we wanted to build a similar maturity framework for data governance programs in an organization. The purpose was to group and distribute each sub-process identified in the chapter 4 model to one maturity level from 2 to 5 as presented in chapter 3.

The method we used to identify which key process areas/sub-processes correspond to which level, was an organized workshop with data governance and enterprise information management consultants from the consulting firm Business & Decision Benelux. The workshop was organized on the 30th of January 2015 at the Business & Decision offices in Brussels. There were a total of 3 participants namely: a data governance expert, a senior enterprise information management consultant and the managing partner for big data and analytics at Business & Decision. The 3 participants received a thorough definition of the concepts they had to use as input: definitions of data governance, CMM as well as what each key process area and sub-process means. Then we presented them with the data governance model from chapter 4 and asked them which elements would correspond to each maturity level of a CMM.

It is important to note that the data governance model identified in chapter 4 encompasses big data governance components as well. During the process of building the model, we have included articles and references pertaining to matters of big data governance as well. During the workshop, we made sure to inform the participants that the expected output was a model which would correspond to both data and big data governance projects.

6.4.5 Data governance process areas by maturity level

Based on the methodology previously described and in line with figure 3.2 in chapter 3, section 3.3.2, figure 6.2 presents a re-worked version of the initial work of O'Regan (2011), this time with the main components of a data governance framework such as they were identified and mapped during the workshop.

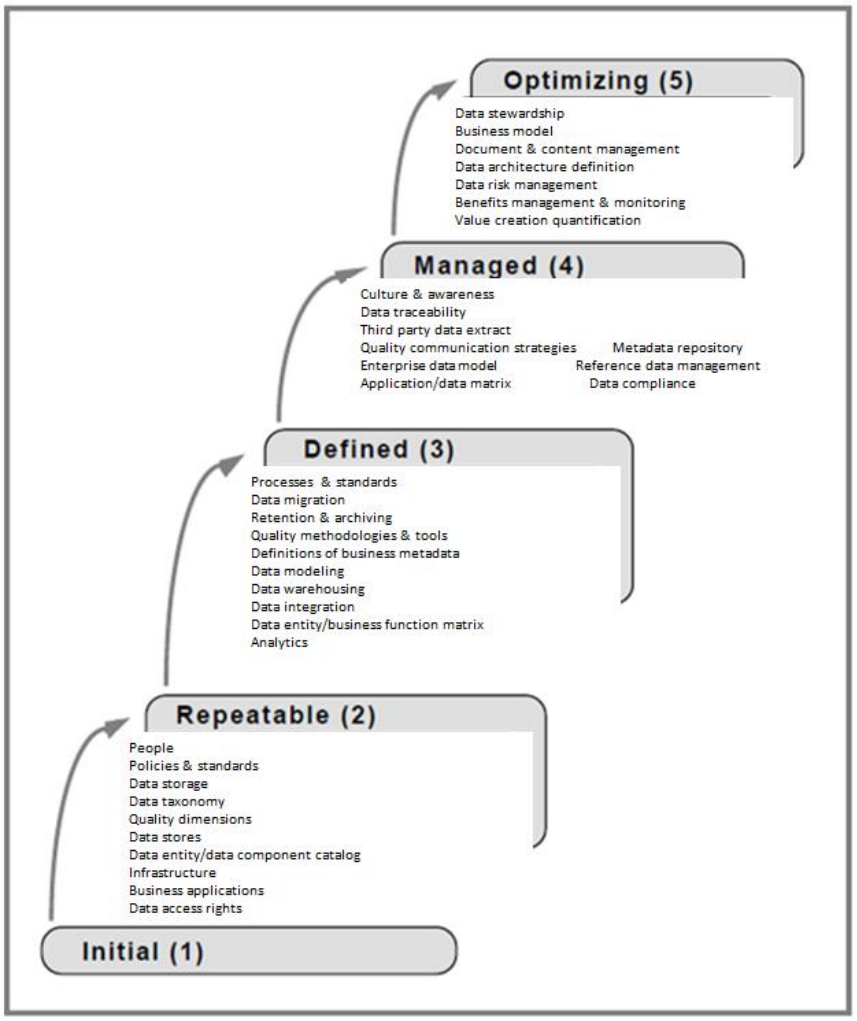

Figure 6.3 The key processes areas of a data governance program by maturity level

Comparing this figure with table 4.2 (Chapter 4, section 4.5.2), we can conclude the following :

- Each maturity level contains at least one sub-process from each key process area identified (expect for level 5 which is the only level where Metrics development and monitoring is mentioned): this is in line with the philosophy of a CMM which suggests a cumulative implementation which builds upon the key process areas in an inferior level of maturity (such as described in Chapter 3);
- Aside from metrics development and monitoring, no key process area is entirely implemented in one level only : this is in line with a staged approach that involves some

basic control objectives first as we move to a higher level of maturity (such as described in Chapter 3);

- Just because one process implementation appears later in a maturity level (such as, for example, Analytics appearing for the first time in level 3) it does not mean that some basic implementation of that process does not yet exist at an inferior level; the difference is that the process is not standardized enterprise-wide at an inferior level and a proper adoption occurs at a later stage.

6.4.6 Basel III implementation model by maturity level

So far, we have presented frameworks specific for data governance in the financial sector as well as industry-free frameworks for data governance in a more general context. The purpose in this sub-section is to bring together the models presented in section 6.4.2 and 6.4.4 in developing a capability maturity model for data (and big data) governance implementations as they have been described in the Basel III "*Principles for effective risk data aggregation and risk reporting*".

We have advanced the idea in section 6.4.2 that Basel III does not imply starting "from scratch" but rather building, improving and enhancing the capabilities a bank already has. In this optic, we will categorize the implementation of the Basel III framework as described in "*Principles for effective risk data aggregation and risk reporting*" to be a level 3 – managed implementation (according to the CMM maturity level description). The rationale is that Basel III builds on the Basel I and II implementations (Basel Committee on Banking Supervision, 2013) which means that there are already some basic, repeatable (level 1 & 2) processes already in place for data governance and risk reporting practices.

With this in mind, we mapped the 2 models described in the previous sections by holding level 3 as the level which best mirrors the Basel III implementation guidelines while for the other levels, we kept the original distribution with small adjustments.

Table 6.3 aggregates the 2 models together and presents the re-arranged version of a capability maturity model for a data and big data governance framework.

Key process area	**Level 1 Initial**	**Level 2 Repeatable**	**Level 3 Defined BASEL III**	**Level 4 Quantitatively managed**	**Level 5 Optimizing**
Roles, structures & policies	Ad-hoc	People Policies & standards	Processes & practices	Culture & awareness	Data stewardship Business model

Data management	Ad-hoc	Data storage Data taxonomy	Data migration Retention & archiving Data traceability Third party data extract Document & content management	-	-	-	-
Data quality management	Ad-hoc	Quality dimensions	Quality methodologies & tools Quality communication strategies	-	-	-	-
Metadata management	Ad-hoc	Definitions of business metadata	Metadata repository	-	-	-	-
Master data management	Ad-hoc	Data stores Data modeling Enterprise data model	Data warehousing Data integration Reference data management	-	-	-	-
Data architecture	Ad-hoc	Data entity/data component catalog	Data entity/business function matrix	Application /data matrix		Data architecture definition	
Technology	Ad-hoc	Infrastructure Business applications	Analytics	-	-	-	-
Security & Privacy	Ad-hoc	Data access rights	Data compliance	Data risk management		-	
Metrics development & monitoring	Ad-hoc	- -	- -	-	-	Benefits management & monitoring Value creation quantification	

Table 6.3 Data and big data governance capability maturity model (under the Basel III implementation)

We can observe that by level 3 – *Defined*, a model implementation following the Basel III guidelines has mainly already implemented most of the data governance process areas we identified in our industry-free governance model. Level 4 and level 5, in this case, constitute an extension of the "foundation" already built by level 3 : level 4 matures the culture of data risk management across the bank as risk management practices are more and more standardized as status quo in daily management of operations while level 5 is concerned with a changing business model that now evaluates just how well its standards and processes are working.

Compared to the original data governance model in figure 6.2 (section 6.4.4), the differences between the distribution of key process areas by level are not that striking. The manner in which Basel III modifies the way a data governance framework should be implemented is by adding a sense of urgency to correctly and timely implementing the safety nets needed for correct risk assessments and reporting. These are data and quality management practices as well as data compliance and analytics.

6.4.7 Empirical testing

We wanted to test how the data governance capability maturity model fits the actual situation in the Belgian financial sector from a data and big data perspective. Given the novelty of the big data projects, we structured an inquiry form which was used to assess

both data and big data governance practices: specifically, we devised a question per sub-process with regards to data governance and one question per sub-process fit to accommodate characteristics for big data projects. We also gave the respondents the possibility of 6 answers on a Likert scale: *I don't know, Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree*. The population we chose were consultants from the researchers network which had previously participated in data governance projects in the Belgian financial sector. The survey was emailed to them and sometimes a follow-up conversation concluded the participation.

We sent out a total of 20 invitations to different profiles: CIO's, CTO's, project managers and consultants which work and/or participate in projects pertaining to the Belgian financial or banking sector. We only received 5 answers in return, which makes for a 25% response rate. These answers came however from 2 banking professionals and 3 consultants. Because the number of answers received was very poor, we would like to draw attention to the fact that we chose to include the empirical results from these questionnaire because of the nature of the work of a consultant: participating in different projects, across different banks, in missions which allow for diversification. This allows the consultant to gain a global view of how the state of data governance practices in banking actually are. We are, however, well aware of the fact that this perceived advantage in the eyes of the researcher, can also be seen as a disadvantage: because of the nature of their work (external employees), consultants might not be exposed to all the mechanisms and workings internal to data governance practices. As such, we present the results of our empirical testing under the reserve that the results offer a very specific, singular view on a broad subject such as data and big data governance and that, as such, these results should not be considered as representative for the entire Belgian financial sector.

6.4.8 Empirical results

Interviewing the different participants to our research revealed that, overall, big data projects, be them short initiatives or more long terms ones, are already being set-up in the banking sector. However, when assessing the maturity level of such big data projects, we find out without surprise, that most of them are in level 1- initial and are as such, composed of ad-hoc, disparate initiatives without any automation or standardization in place. We found, however, that there exists a strong culture & awareness for big data projects among the participants as most decision-making practices involve taking all potential available data in consideration. Such a data-oriented mindset is not sustained by accurate standards and procedures.

Overall, thus for both data and big data governance oriented practices, the results presented in figure 6.4 aggregate the computed maturity levels for each sub-process and gives a global overview of how banks are performing for each key process area, compared to the Basel III requirements.

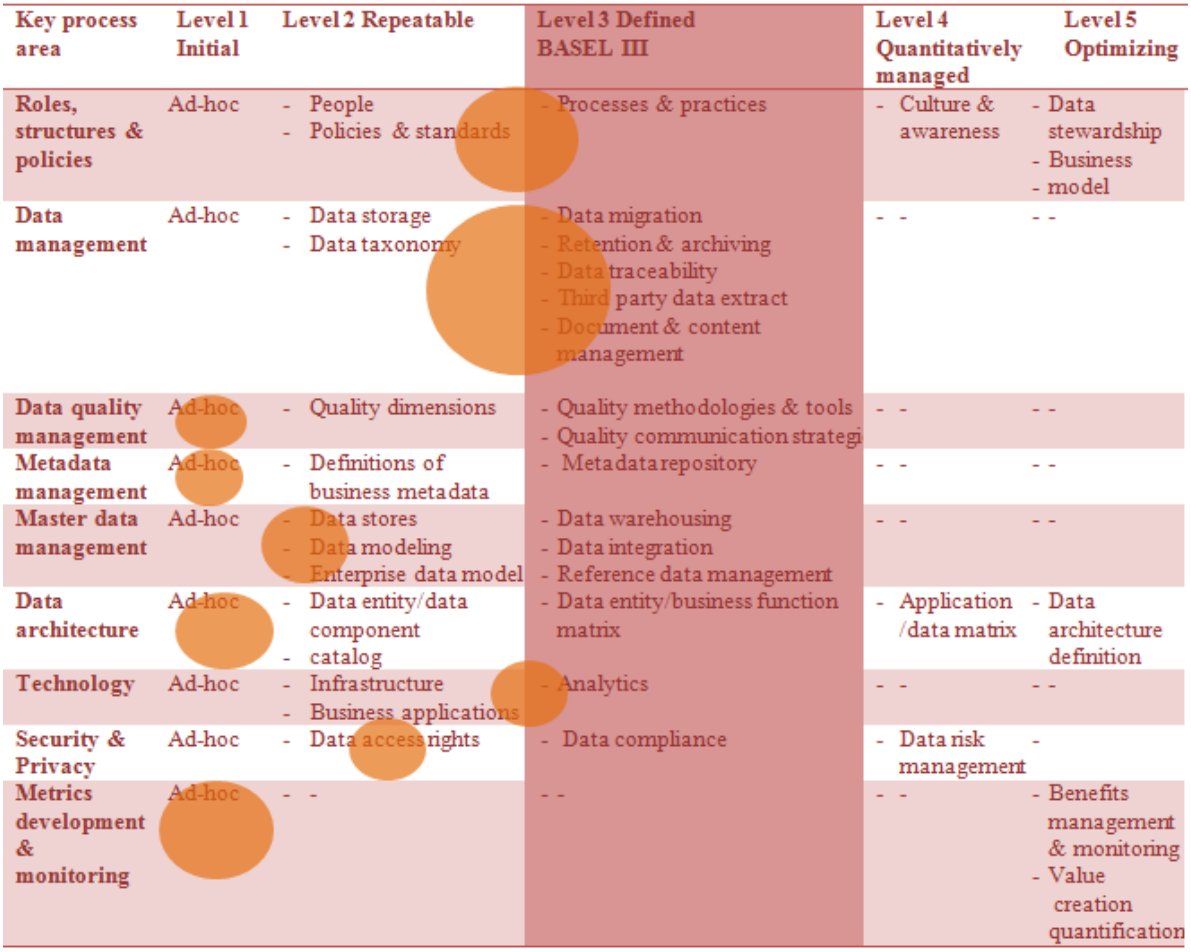

Figure 6.4 Performance of the Belgian financial sector in data governance practices

The orange bubbles represent the observed maturity level. Their position gives an indication that a key process area might be situated between levels such as for example data management practices which have been evaluated between level 2 and level 3 of maturity. We have analyzed the observed performance versus the desired, industry-level (in this case Basel III) performance (level 3).

Overall, we notice that while policies and standards exist in place to support operations, these are yet to have been transformed in processes and practices to sustain a thorough implementation of the Basel III principles and guidelines.

From a data management perspective, it seems that banks have already implemented many of the sub-processes needed for solid data collection across the different business units. However, when it comes to actual integration, aggregation and centralization of

the reference data, their practices are still mostly ad-hoc and lacking a repeatability which allows for statistical measuring. This is not surprising seeing that data architecture practices are evaluated at level 1 : no solid architecture in place to support a sustainable concentration and use of available data across the bank results in a scattered data usage. It is however contradictory because infrastructure is evaluated between level 2 and 3 which means that the technology needed to sustain the "*consumption*" of data across the bank exists and analytic activities are being performed nonetheless. This can be the symptom of having purchased an array of technologies which have yet to be integrated with each other and which allow each business unit to perform its own analysis and modeling activities without much consideration as to how this data can be enriched or enhanced with other available sources. It is even more unexpected since data access rights are being evaluated at level 2 which means that data stored in organizational silos can be made available on request for the users interested in using it.

As far as metrics development and monitoring, banks have yet to develop the indicators needed to assess how well their processes and practices perform against the baseline.

6.5 Chapter conclusions

Data governance and big data governance are important topics in any industry, sector or organization. Indeed, far from only allowing for the definition of policies and roles pertaining to the *what* and *who*, their purpose extends to building the staging area of using data as an asset across the enterprise.

This subject, being quite a generalist one in literature, has allowed us to build a data governance model which can encompass the more "*traditional, small*" data as well as new emerging trends such as big data sources. We have shown that, with big data governance, it is hard to make the distinction between where data governance stops and big data governance begins. This difficulty stems mostly from the fact that big data governance builds upon the bricks of data governance : without a solid data foundation, no big data will ever aspire to become "*big enough*".

When analyzing such data and big data frameworks in the financial sectors, the challenge augments in scope and intricacy because of complex regulations, compliance measures and general size and importance of such a sector. We tried to refrain from building a model which unrealistically fails to reflect the actual conditions in the banking sector. Building upon the Basel III framework seemed the best solution to balance the practical needs of the financial sector with the theoretical precision of academic modeling. In this context, the capability maturity model presented in this chapter offers

the possibility to translate data governance practices from all sectors to one sector as the model created is flexible enough to be tailored and customized to requirements.

The empirical data gathered, while far from being significant, points out to the shortcomings and level of maturity of current data and big data governance practices in the outlook of the Basel III implementation. The results we gathered are just an indication of the milestones yet to achieve before reaching the compliant level such as indicated by the Basel Committee on Banking Supervision.

The results presented in this chapter allow however to draw a potential roadmap for improvement and to build on top of the bricks of data governance programs already existing in the financial sector today.

7. CONCLUSIONS

Governance is an importance concept in organizational studies because it bridges a strategy to an actual implementation roadmap: concrete guidelines can only stem from existing policies and practices. In today's business landscape, governance has become linked with IT practices because of IT omnipresence in the enterprise: as more and more regulations request more transparency in business operations, IT is seen as the means of complying with the demands. We showed how IT governance reposes on decision-making structures which enable its positioning as a strategic partner in the enterprise: problem identification and problem solution structures which enable IT to scan for potential problems before they occur and implement the necessary safety nets to stop them from occurring. IT is no longer positioned as the "*back office*" of an organization where service level agreements and procedural components "*rule*", but rather an important contributor and "*supplier*" of value across the business.

Properly using an IT governance framework means understanding what drives maturity, performance and capability, the characteristics needed to confirm the evolution to a superior model of performance. We then turn to the capability maturity models which allow us to build a roadmap for any organization wanting to embark on process improvement activities. Because of their volatility and level-like distribution, capability maturity models can be transferred to any domain and customized enough to allow for building a reference map to guide improvement, enhancement and progress of current processes.

Data governance comes to complete or complement IT governance as the opinions in literature are divergent as to what comprises who and how the differentiation between the two must be done. As the same theories are used to explain both concepts and as the relationship between IT and data governance cannot be underestimated, we advise to treat both subjects together while acknowledging the differences that make them separate entities in a business. IT governance, with its strategically positioning in an enterprise, builds the backbone of technology and infrastructure to support operations. Data governance is, in this optic, the lifeblood that "*pumps*" this backbone into further development. Data, as an enterprise-wide asset should rest under the umbrella of IT governance and be protected as such by the best infrastructure which allows for efficient exploitation. Data, in the multiple facets and forms it comes in, can provide valuable information to both IT and the business, the distinction between the two is wrongly made in practice: data should be the missing link which connects IT and the business. As

a by-product of any organization, data is valuable even when data is not the main product an organizations sells. This reflection only is enough to "*reconcile*" the views which position data and IT separately.

How does this differentiation apply when it comes to the big data emerging trend? In our view, the reconciliation between the two is even more important as big data infrastructure and data analysis needs should be supported by enhancing the current IT capabilities. Building separate or new IT structures seems unreasonable if the current IT system can be scaled to sustain new computational needs. Of course, the challenge resides again in the relationship between the business use of big data and the IT capabilities needed to manage this data. We stress the importance of early defining the objectives needed to resolve the potential conflicts which may appear between how this data will be managed and what it will be used for.

These challenges and conflicts can be found in the financial sector as big data related issues require more and more attention in order to keep a competitive advantage in an ever-competing environment. In this environment, data governance issues are of crucial importance and not only because of the potential value delivered by the data collected but because the data collected ensures the flow of operations suffers no shocks. Ensuring that 1) daily transactions and operations occur smoothly, 2) the data collected is used at its correct value and 3) risk can be early identified by correctly analyzing data patterns, could not be envisaged without a solid data governance structure across the financial sector. Unfortunately, as we have shown, this is yet to be the industry standard as banks move slowly across regulations and compliance rules to set up standardized procedures which maximize data usage. Data has nowadays the potential to become the early warning system to dangerous imbalances in the system, if used correctly and sustained by strong governance practices.

The practical contribution of our work is the attempt to provide such a strong data governance model to the financial sector. It is easy to talk about what governance is or what governance should do but it is hard to understand and link all of the potential elements which sustain it. In this paper, we tried to uncover these elements by analyzing the work which has already been done in this domain. We brought together the disparate elements of such models and homogenously standardized them to fit the specificities of any sector. We chose to apply it to the financial sector because of its intrinsic need for standards and frameworks: the notion of statistical control mentioned by the capability maturity models states that a practice which is under statistical control will always produce the same results. Regulations and compliance charters try to normalize practices across the financial sector in order to ensure that good practices are

synchronized across the industry while avoiding that bad practices impact the equilibrium of the whole.

The model we presented in our paper is a model which follows the structure of the Basel III regulation framework while fitting the elements characteristic to a particular sector. It also takes into the account that while regulation ensures for uniformity in practices across the financial sector, data governance capabilities offer a potential competitive advantage which should not be underestimated. How does this competitive advantage fit into our model without *"leveling the ground"* too much? While we developed the model with the idea in mind that it is easier to assess and control risk reporting practices across an industry if every single participant is using the same measures of risk, it is possible for each entity to keep its competitive advantage by playing on the way the framework is implemented. The model we have developed is tool-free and can be implemented by using the frameworks, technologies and design patterns the user choses as long as the rules of implementation are respected. The choice of technology constitutes, in this case, the source of competitive advantage.

This model is, as we have shown, not a static one and while implementing its levels such as they were defined in the maturity structures, we have to consider that an early implementation (such as for example, policies and standards implemented in level 2) has to be flexible enough to allow for enhancements and scalability during later levels. One must also take into consideration the misconception that the implementation of a specific, standard framework hinders innovation and suffocates any improvement initiatives. We would like to draw the attention to the remark made earlier in chapter 6 by Flood et al.(2012), which stated that innovations are highly regarded and encouraged if there exists a synchronization between the front and back-office. With his view in mind, improvements can be made to the model we developed and these improvements can very well be in line with the Basel III risk reporting principles but one must ensure that the proper infrastructure and policies exist in place to support the safe propagation of the innovation across the value-chain.

It is not our interest to advance a rigid model which obstructs the flexibility needed when dealing with governance practices: while we acknowledge how important a strong pillar is to sustain an architecture, we must however allow "the architect" to deploy its creativity and imagination when designing the final edifice. In the context of our work, this remains the main challenge faced by the financial sector: building the pillar for data governance practices while allowing innovation to propagate through its frontiers.

Bibliography

Afzali, P., Azmayandeh, E., Nassiri, R., & Latif Shabgahi, G. (2010). Effective governance through simultaneous use of COBIT and Val IT. *2010 International Conference on Education and Management Technology*, 46-50.

Aiken, P., David Allen , M., Parker, B., & Mattia, A. (2007). Measuring data management practice maturity. *Computer*, 40(4), 42-50.

Alagha, H.(2013).Examining the Relationship between IT Governance Domains, Maturity, Mechanisms, and Performance: An Empirical Study toward a Conceptual Framework. *Tenth International Conference on Information Technology: New Generations (ITNG)*, 767-772.

Ali-ud-din Khan, M., Uddin, M.F., & Gupta, N.(2014). Seven V's of Big Data understanding Big Data to extract value. *2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1)*, 1-5.

Alves de Freitas, P., dos Reis, E.,A., Michel, S.W., Gronovicz, M.,E., & Rodrigues, M.,A.,M. (2013). Information governance, big data and data quality. *IEEE 16th International Conference on Computational Science and Engineering*, 1142-1143.

Apache Hive. (2015). *Apache Hive*. Retrieved from https://cwiki.apache.org/confluence/display/Hive/Home

Apache Mahout. (2014). *What is Apache Mahout* ? Retrieved from https://mahout.apache.org/

ARMA International.(2015). *The generally accepted recordkeeping principles*. Retrieved from http://www.arma.org/r2/generally-accepted-br-recordkeeping-principles

Bahjat El-Darwiche, B., Koch,V., Meer, D. Shehadi, R., & Tohme, W. (2014). Big data maturity: An action plan for policymakers and executives. *The Global Information Technology Report, World Economic Forum,* 43-51.

Barnard, K., & Avery, A. (2011). *Basel III v Dodd-Frank : what does it mean for US banks*. Retrieved from http://whoswholegal.com/news/features/article/28829/basel-iii-v-dodd-frank-does-mean-us-banks/

Basel Committee on Banking Supervision. (2013). Principles for effective risk data aggregation and risk reporting. *Bank for International Settlements*.

Bedi, P., Jindal, V., & Gautam, A. (2014). Beginning with big data simplified. *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, 1-7.

Bohlouli, M., Schulz, F., Angelis, L., & Pahor, D. (2013). Towards an integrated platform for Big Data analytics. In Bohlouli, M., Schulz, F., Angelis, L., PAhor, D., Brandic, I., Atlan, D., & Tate, R. (Eds.), *Integration of practice-oriented Knowledge Technology: Trends and prospectives* (pp. 48-55). Berlin: Springer Berlin Heidelberg.

Borthakur, D. (2012). *The Hadoop Distributed File System : architecture and design*. Retrieved from http://hadoop.apache.org/docs/r0.18.0/hdfs_design.pdf

Brammertz, W. (2012). The office of Financial Research and Operational Risk. In V. Lemieux (Eds.), *Financial Analysis and risk management: data governance, analytics and life cycle management* (pp. 47-73). Frankfurt: Springer Berlin Heidelberg.

Buhl, HU., Röglinger, M., Moser, F., & Heidemann J. (2013). Big Data : A Fashionable Topic with(out) Sustainable Relevance for Research and Practice? *Business & Information Systems Engineering*, 5(2), 65-69.

Chapple, M.(2013). Speaking the same language: building a data governance program for institutional impact. *EDUCAUSE Review*, 48(6), 14-27.

Chen, M., Mao,S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and applications*, 19(2), 171-209.

Cheong, L.,K., & Chang, V.(2007). The need for data governance : a case study. *18th Australasian Conference on Information Systems*, 999-1008.

Chrissis, M. B., Konrad, M., & Shrum, S. (2003). *CMMI Guidelines for Process Integration and Product Improvement*. Addison-Wesley Longman Publishing Co.

Chukwa. (2015). *Chukwa: a large-scale monitoring system*. Retrieved from http://wiki.apache.org/hadoop/Chukwa?action=AttachFile&do=view&target=chukwa_cca08.pdf

Cloudera. (2014). *Flume Cookbook*. Retrieved from http://archive.cloudera.com/cdh/3/flume/Cookbook/index.html#_introduction

Curley, M. (2008). Introducing an IT capability maturity framework. *Enterprise Information systems: lecture notes in Business Information processing*, 12, 63-78.

Datskovsky, G. (2010). Information Governance. In Lamm J. , Blount S., Boston S., Camm M., Cirabisi R., E. Cooper N., Fox C., V. Handal K., E. McCracken W., Meyer J., Scheil H., Srulowitz A., & Zanella R. (Eds), *Under Control* (157-181). Apress.

De Abreu Faria, F., Macada, AC.G., Kumar, K. (2013) . Information governance in the banking industry. *46th Hawaii International Conference on System Sciences (HICSS)*, 4436-4445.

De Haes, S., Van Grembergen, W.(2005). IT Governance structures, processes and relational mechanisms: achieving IT/business alignment in a major Belgian financial group. *Proceedings of the 38th Hawaii International Conference on System Sciences*, 1-10.

Demchenko, Y., de Laat, C., Membrey, P.(2014). Defining architecture components of the Big Data Ecosystem. *2014 International Conference on Collaboration Technologies and Systems (CTS)*,104-112.

Dyché, J.(2011). A data governance primer. *Baselinemag*. 28-29.

Ebner, K., Buhnen, T., & Urbach, N.(2014).Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments. *47th Hawaii International Conference on System Sciences (HICSS)*, 3748-3757.

Esmaili, H., B., Gardesh, H.,Shadrokh Sikari, S.(2010). Strategic alignment: ITIL perspective. *2nd International Conference on Computer Technology and Development (ICCTD 2010)*, 550-555.

Facebook. (2008). *Scribe makes its open source debut*. Retrieved from https://www.facebook.com/notes/facebook/scribe-makes-its-open-source-debut/35548642130

Fernandes, L., O'Connor, M. (2009). Data governance and data stewardship – Critical issues in the move towards EHRs and HIE. *Journal of AHIMA (American Health Information Management Association)*, 80(5), 9-36.

Finance Lab. (2014). *Persoonlijke benadering klanten via big data bij ING*. Retrieved from http://financelabblog.wordpress.com/2014/03/10/persoonlijke-benadering-klanten-via-big-data-bij-ing/

Financial Stability Board and the International Monetary Fund. (2009) . *The financial crisis and information gaps: report to the G-20 finance ministers and central bank governors*. Retrieved from http://www.financialstabilityboard.org/publications/r_091107e.pdf

Flood, M., Mendelowitz, A., & Nichols, B. (2012). Monitoring Financial stability in a complex world. In V. Lemieux (Ed.), *Financial Analysis and risk management: data governance, analytics and life cycle management* (pp. 15-47). Frankfurt: Springer Berlin Heidelberg.

Georges, G., Haas, M., & Pentland, A. (2014). Big data and management. *Academy of Management Journal*, 52(2), 321-326.

Ghemawat, S., Gobioff, H., & Leung, S-T. (2003). The Google File System. *Proceedings of the nineteenth ACM Symposium on Operating Systems Principles*, 29-.

Griffin, J.(2010a). Four critical principles of data governance success. *Information Management Journal*, 29-30.

Griffin, J.(2010b). Implementing a data governance initiative. *Information Management Journal*, 27-28.

Griffin, J.(2005). Data governance: a strategy for success, part 2. *DM Review*, 15(8), 15-.

Griffin, J.(2008). Data governance: the key to enterprise data management. *DM Review*, 27-.

Hagmann, J.(2013). Information governance- beyond the buzz. *Records Management*.

Hansmann, T., & Niemeyer, P.(2014). Big Data - Characterizing an Emerging Research Field Using Topic Models. *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT),* 1, 43-51.

Harris, J. (2012). *Data Governance Frameworks are like Jigsaw Puzzles*. Retrieved from http://www.ocdqblog.com/home/data-governance-frameworks-are-like-jigsaw-puzzles.html?rq=Scott%20Berkun

Hay, J. (2014). Data governance gamification. *Business Intelligence Journal*, 19(1), 30-35.

Heier, H., Borgman, H.P., & Mileos, C. (2009). Examining the Relationship between IT Governance Software, Processes, and Business Value: A Quantitative Research Approach. *Proceedings of the 42nd Hawaii International Conference on System Sciences*, 1-11.

Henserson, J.C, & Venatraman, N. (1993). Strategic alignment : leveraging Information Technology for transforming organizations. *IBM Systems Journal*, 32(1), 472-484.

Hu, H., Wen, Y., Chua, T-S., Li,X. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. Access, *IEEE*, 2, 652-687.

Humphrey, W.S. (1988). Characterizing the software process: a maturity framework. Software, *IEEE*, 5(2), 73-79.

IBM (2014). *The 5 game changing big data use cases*. Retrieved from http://www-01.ibm.com/software/data/bigdata/use-cases.html

IBM. (2007). Building a roadmap for effective data governance. *IBM Data Governance Council Maturity Model*.

IBM. (2014) *Information Governance Principles and practices for Big Data landscape*. Retrieved from http://www.redbooks.ibm.com/abstracts/sg248165.html?Open

Isaca. (2012). *COBIT 5*. Retrieved from http://www.isaca.org/chapters2/New-York-Metropolitan/membership/Documents/2012-04-30%20Spring%20Conference-Meeting/3%20Barnier%20VBA%20COBIT5.pdf

Katal, A., Wazid, M., & Goudar, R.H.(2013).Big data: Issues, challenges, tools and Good practices. *Sixth International Conference on Contemporary Computing (IC3)*, 404-409.

Khatri, V., & V. Brown, C. (2010). Designing Data Governance. *Communications of the ACM*, 53(1), 148.

Kindler, T. (2013). *The road to Basel III: how financial institutions can meet new data-management challenges*. USA : Teradata Corporation.

Krishnan, K.(2013). *Data warehousing in the age of big data*. Chicago, Illinois: Morgan Kaufmann.

Kuruzovich, J., Bassellier, G., & Sambamurthy, V. (2012). IT Governance Processes and IT Alignment: Viewpoints from the Board of Directors. *45th Hawaii International Conference on System Science (HICSS)*, 5043-5052.

Ladley, J. (2012). *Data Governance : How to Design, Deploy, and Sustain an Effective Data Governance Program*. USA : Elsevier.

Lahrmann, G., Marx, F., Winter, R., & Wortmann, F. (2011). Business Intelligence Maturity: Development and Evaluation of a Theoretical Model. System Sciences (HICSS), *44th Hawaii International Conference on System Sciences (HICSS)*,1-10.

Lemieux, V. (2012*). Financial Analysis and Risk Management: Data Governance, Analytics and Life Cycle Management*. Frankfurt: Springer Berlin Heidelberg.

Lewis, E., Millar, G. (2009). The Viable Governance Model - A Theoretical Model for the Governance of IT. *42nd Hawaii International Conference on System Sciences*, 1-10.

Lucas, A. (2011). Corporate Data Quality Management: Towards a Meta-Framework. *International Conference on Management and Service Science (MASS)*, 1-6.

Maes, R.(1999). A Generic Framework for Information Management. *Primavera Working Paper.*

McGilvray, D.(2007). Data governance: a necessity in an integrated information world. *DM Review*, 16(12), 25-30.

McKinsey (2014). *Presentation: Big Data and advanced analytics: 16 use cases*. Retrieved from http://mckinseyonmarketingandsales.com/presentation-big-data-and-advanced-analytics-16-use-cases

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). *Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics (1st ed.)*. CA, USA: Apress, Berkely.

Morabito, V. (2014). *Trends and challenges in digital business innovation*. Switzerland: Springer International Publishing Switzerland.

Mosley, M. (2008). *DAMA-DMBOK Functional Framework Version 3*. Retrieved from http://www.dama.org/files/public/dama-dmbok_functional_framework_v3_02_20080910.pdf

Nassiri R., Ghayekhloo, S., & Shabgahi, G.L. (2009). A novel approach for IT governance : a practitioner view. *International Conference on Computer Technology and Development*, 217-221.

O'Regan, G. (2011). *Introduction to software process improvement*. London: Springer-Verlag Limited.

Ohata, M., & Kumar, A. (2012). Big Data : A boon to business intelligence. *Financial Executive*, 28(7), 63.

Otto, B. (2011). Data Governance. *Business & Information Systems Engineering*, 3(4), 241-.

Paulk, M.,C. (2009). A history of the capability maturity model for software. *ASQ Software Quality Professional*, 12(1), 5-19.

Paulk, M.,C., Curtis, B., Chrissis, M.,B., & Weber, C. (1993). Capability maturity model for software ver. 1.1. *Software Engineering Institute CMU/SEI '93-TR*.

Ploder, C., & Fink, K. (2008). Decision Support Framework for the implementation of IT-Governance. *Proceedings of the 41st Hawaii International Conference on System Sciences*, 1-10.

Rajpurohit, A.(2013).Big data for business managers — Bridging the gap between potential and value. *IEEE International Conference on Big Data*, 29-31.

Ribbers, P.M.A., Peterson, R.R., & Parker, M.M. (2002). *Proceedings of the 35th Hawaii International Conference on System Sciences*,1-12.

Simonsson, M.,& Ekstedt, M. (2006). Getting the priorities right: literature vs practice on IT governance. *PICMET 2006 Proceedings*, 18-26.

Skinner, T., H. (2015). *Does Basel III apply to the community bank ?* USA: SAS Institute.

Soares, S.(2011). *Selling Information Governance to the Business*. MC Press, Ketchum, ID.

Sucha, M. (2014). Beyond the hype: Data management and data governance. *Feliciter (Canadian Library Association),* 60(2), 26-29.

Tallon, P.P. (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *Computer* , 46(6), 32-38.

Tamasauska, D., Liutvinavicius, M., Sakalauskas, V., & Kriksciuniene, D. (2013). Research of conventional data mining tools for Big Data handling in finance institutions. *Business Information Processing*, 160, 35-46.

Team, S. C. P. (2010). *CMMI for Development v1. 3*. Lulu. com.

Tekiner, F., & Keane, J.A.(2013). Big Data Framework. *IEEE International Conference on Systems, Man, and Cybernetics (SMC),* 1494-1499.

Thamir, A., & Theodoulidis, B. (2013). Business intelligence maturity models: information management perspective. *Communications in Computer and Information Science*, 403, 198-221.

Todd, G. (2008). Data Governance: the enabler of high performance. *DM Review*, 18(5), 30.

TOGAF. (2015). *Phase C: Information Systems Architectures - Data Architecture*. Retrieved from http://pubs.opengroup.org/architecture/togaf8-doc/arch/chap08.html

U.S. Congress (2010). *H.R.4173 - Dodd-Frank Wall Street Reform and Consumer Protection Act*. Retrieved from https://www.congress.gov/bill/111th-congress/house-bill/4173?q=%7B%22search%22%3A%5B%22Dodd-Frank+wall+street+reform+and+consumer+protection+act%22%5D%7D

Van Grembergen, W., De Haes, S., & Guldentops, E. (2004). Structures, Processes and Relational Mechanisms for IT Governance. In W. Van Grembergen (Ed.), *Strategies for Information Technology Governance* (pp. 1-36). Hershey, PA: Idea Group Publishing.

Van Grembergen, W., De Haes, S., & Guldentops, E. (n.d). *Structures, Processes and Relational Mechanisms for IT Governance: theories and practices*. Universiteit Antwerpt Management School. Retrieved from http://www.antwerpmanagementschool.be/media/287503/IT%20Gov%20theories%20and%20practices.pdf

Van Grembergen, W., De Haes, S.(2009) *Enterprise Governance of Information Technology*. Springer. New York

Van Leemputten, P. (2014). *KBC investeert half miljard euro in big data*. Datanews. Retrieved from http://datanews.knack.be/ict/nieuws/kbc-investeert-half-miljard-euro-in-big-data/article-4000662817321.html

Vesset, D., Morris, H.D., Little, G., Borovick, L., Feldman, S., Eastwood, M., Woo, B., Villars, R.L., …, Yezhkova, N. (2012). Worldwide Big Data technology and services 2012-2015 forecast. *IDC*, 233485(1).

Waddington, D.(2008). Adoption of data governance by business. *DM Review*, 18(12), 32.

Webb, P., Pollard, C., & Ridley G. (2006). Attempting to define IT governance: wisdom or folly? *Proceedings of the 39th Hawaii International Conference on System Sciences*, 1-10.

Weber, C. V., Curtis, B., & Chrissis, M. B. (1994). *The capability maturity model: Guidelines for improving the software process* (Vol. 441). Reading, MA: Addison-Wesley.

Weber, K., Otto, B., & Osterle, H. (2009). One size does not fit all—a contingency approach to data governance. *Journal of Data and Information Quality*, 1(1), 4:2-4:27

Wielki, J.(2013).Implementation of the Big Data concept in organizations - possibilities, impediments and challenges. *Federated Conference on Computer Science and Information Systems*, 985-989.

Zhang, J., Chen, Y., & Li, T. (2013). Opportunities of innovation under challenges of big data. *10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 669-673

Appendices

Appendix A

Brief description of the different key process areas per level (O'Regan, 2011, pp.56-58)

Maturity level	Process area	Description of process area
Level 2	REQM	*Requirements management* This process area is concerned with managing the requirements for the project and ensuring that the requirements, project plan(s), and work products are kept consistent
	PP	*Project planning* This process area is concerned with estimation for the project, developing and obtaining commitment to the project plan, and maintaining the plan
	PMC	*Project monitoring and control* This process area is concerned with monitoring progress with the project and taking corrective action when project performance deviates from the plan
	SAM	*Supplier agreement management* This process area is concerned with the selection of suppliers, documenting the (legal) agreement/statement of work with the supplier, and managing the supplier during the execution of the agreement
	MA	*Measurement and analysis* This process area is concerned with determining management information needs and measurement objectives. Measures are then specified to meet these objectives, and data collection and analysis procedures are defined. Data are collected and measurements analysed and communicated
	PPQA	*Process and product quality assurance* This process area is concerned with providing objective visibility to management on the extent of process compliance. Non-compliance issues are documented and resolved by the project team
	CM	*Configuration management* This process area is concerned with the management of change. It involves setting up a configuration management system; identifying the items that will be subject to change control and controlling changes to them. Configuration audits are conducted

Maturity level	Process area	Description of process area
Level 3	RD	*Requirements development* This process area is concerned with eliciting and defining customer, product, and product–component requirements and analysing and validating the requirements
	TS	*Technical solution* This process area is concerned with the design, development, and implementation of an appropriate solution to the customer requirements
	PI	*Product integration* This process area is concerned with the assembly of the product components to deliver the product, and verifying that the assembled components function correctly together
	VER	*Verification* This process area is concerned with ensuring that selected work products satisfy their specified requirements. This is achieved by peer reviews and testing
	VAL	*Validation* This process area is concerned with demonstrating that the product or product component is fit for purpose and satisfies its intended use
	OPF	*Organization process focus* This process area is concerned with planning and implementing process improvements based on a clear understanding of the current strengths and weakness of the organization's processes
	OPF	*Organization process definition* This process area is concerned with creating and maintaining a usable set of organization processes. This allows consistent process performance across the organization
	OT	*Organization training* This process area is concerned with developing the skills and knowledge of people to enable them to perform their roles effectively
	IPM	*Integrated project management* This process area is concerned with tailoring the organization set of standard processes to define the project's defined process. The project is managed according to the project's defined process
	RSKM	*Risk management* This process area is concerned with identifying risks and determining their probability of occurrence and impact should they occur. Risks are identified and managed throughout the project
	DAR	*Decision analysis and resolution* This process area is concerned with formal decision-making. It involves identifying options, specifying evaluation criteria and method, performing the evaluation, and recommending a solution

Maturity level	Process area	Description of process area
Level 4	OPP	*Organization process performance* This process area is concerned with obtaining a quantitative understanding of the performance of selected organization processes in order to quantitatively manage projects in the organization
	QPM	*Quantitative project management* This process area is concerned with quantitatively managing the project's defined process to achieve the project's quality and performance objectives
Level 5	OID	*Organization innovation and deployment* This process area is concerned with incremental and innovative process improvements
	QPM	*Causal analysis and resolution* This process area is concerned with identifying causes of defects and taking corrective action to prevent a re-occurrence in the future

Appendix B

Mapping of key process areas to sources

All references pertaining to elements which were identified as being part of data governance, information governance, data management or information management programs have been categorized under the "Key process area" label. The "Source" label indicates the reference work. The table includes the sources for each process as it was initially identified.

KEY PROCESS AREA	SOURCE
Compliance	IBM (2014), Chapple (2013)
Data compliance	Todd (2008)
Regulations & Compliance	IBM (2007)
Data architecture	Lucas (2011), IBM (2014), Diché (2011), Griffin (2010), Hay (2014)
Data architecture management	DAMA-DMBOK (2008)
Enterprise Architecture	IBM (2007)
Data Management	Demchenko, De Laat, Membrey (2014)
Data development	DAMA-DMBOK (2008)
Data development	Aiken et al.(2007)
Data management	Diché (2011)
Retention & archiving	Chapple (2013)
Document & Content Management	DAMA-DMBOK (2008)
Data taxonomy	Todd (2008)
Data traceability	Todd (2008)
Data requirements	Diché (2011)
Data content	Hay (2014)
Data administration	Diché (2011)
Data archiving	Todd (2008)

Data migration	Todd (2008)
Audit Information Logging & Reporting	IBM (2014)
Information management & usage	IBM (2007)
Data storage	Todd (2008)
Third Party Data extract	Cheong & Chang (2007)
Data profiling	Cheong & Chang (2007), Todd (2008)
Data profiling tool	Cheong & Chang (2007)
Data quality	Khatri & Brown (2010), Diché (2011)
Data quality communication strategies	Lucas (2011)
Data quality dimensions	Lucas (2011)
Data Quality Management	DAMA-DMBOK (2008), IBM (2014), Lucas (2011)
Data quality methodologies, technologies & tools	Lucas (2011)
Data cleansing (data cleansing tool)	Cheong & Chang (2007)
Data cleansing	Todd (2008)
Quality & consistency	Chapple (2013)
Data Stewardship	Aiken et al.(2007), Diché (2011), Todd (2008), IBM (2014)
Data ownership	Todd (2008)
Data custodianship	Cheong & Chang (2007)
Governance metrics	Griffin (2010)
Metrics	Griffin (2011)
Metrics development & monitoring	Cheong & Chang (2007)
Benefits management & reporting	De Haes & Van Grembergen (2005)
Value Creation	IBM (2014)
Data monitoring	Todd (2008)
Information Life-cycle management	IBM (2014), IBM (2007)
Data retention	Todd (2008)
Data retirement	Todd (2008)
Data lifecycle	Khatri & Brown (2010)
Master data management	DAMA-DMBOK (2008), Todd (2008)
Reference data management	Todd (2008)
Enterprise Data Model	IBM (2007)
Enterprise Data Stores	IBM (2007)
Data Warehousing	DAMA-DMBOK (2008)
Data model/types	Demchenko, De Laat, Membrey (2014)
Data integration	Lucas (2011), Aiken et al.(2007)
Data modeling	Todd (2008)
Meta data management	DAMA-DMBOK (2008), Diché (2011), Cheong & Chang (2007), Todd (2008)
Metadata	IBM (2014)
Metadata	Khatri & Brown (2010)
Metadata repository	Cheong & Chang (2007)
Business metadata	Hay (2014)
Definitions of business metadata	Hay (2014)
Organizational bodies & policies	Cheong & Chang (2007)
Organization & policies	Griffin (2010)

Organizational structures (& Culture)	IBM (2007)
Organizational structures (& awareness)	IBM (2014)
People	De Abreu Faria, Maçada, Kumar (2013)
Policies & standards	Chapple (2013)
Policies (& practices)	De Abreu Faria, Maçada, Kumar (2013)
Policy	IBM (2014)
Principles & standards	Griffin (2010)
Processes& practices	Griffin (2010)
Sponsorship	Lucas (2011)
Governance policies	Dyché (2011)
Governance Structure	Cheong & Chang (2007)
Roles, responsibilities & requirements	Griffin (2011)
Roadmap	Griffin (2011)
Executive Sponsorship	Diché (2011)
Decision rights	Cheong & Chang (2007)
Data program coordination	Aiken et al.(2007)
Data governance structure	Cheong & Chang (2007)
Data asset use	Aiken et al.(2007)
Business Model	Mohanty, Jagadeesh, Srivatsa (2013)
Issue escalation process	Cheong & Chang (2007)
Data policies	Todd (2008)
Data policy	Lucas (2011)
Data principles	Khatri & Brown (2010)
Data standards	Todd (2008)
User group charter	Cheong & Chang (2007)
Security	Demchenko, De Laat, Membrey (2014), IBM (2014)
Security & Access rights	Dyché (2011)
Security & Privacy	Chapple (2013)
Data privacy	Todd (2008)
Data access	Khatri & Brown (2010), Todd (2008)
Data Risk Management	IBM (2014)
Data security management	DAMA-DMBOK (2008)
Data access	Chapple (2013)
Technology	IBM (2007), Cheong & Chang (2007), De Abreu Faria, Maçada, Kumar (2013), Griffin (2010), Chapple (2013)
Infrastructure	Demchenko, De Laat, Membrey (2014)
Analytics	Demchenko, De Laat, Membrey (2014)
Business Applications	IBM (2007)
Data support operations	Aiken et al.(2007)
Database operations management	DAMA-DMBOK (2008)

Appendix C

Mapping of key process areas to sources : frequency

All references pertaining to elements which were identified as being part of data governance, information governance, data management or information management programs have been categorized under the "Key process area" label. The "Source" label indicates the reference work. "Count" provides an aggregated count of the number of times a process has been mentioned by more than one source e.g : Compliance has been mentioned by 4 different sources, data access has been mentioned by 3 different sources

KEY PROCESS AREA	SOURCE	COUNT
Compliance	IBM (2014)	4
Compliance	Chapple (2013)	4
Data compliance	Todd (2008)	4
Regulations & Compliance	IBM (2007)	4
Data access	Khatri & Brown (2010)	3
Data access	Todd (2008)	3
Data access	Chapple (2013)	3
Data architecture	Lucas (2011)	7
Data Architecture	IBM (2014)	7
Data architecture	Diché (2011)	7
Data architecture	Griffin (2010)	7
Data architecture	Hay (2014)	7
Data architecture management	DAMA-DMBOK (2008)	7
Enterprise Architecture	IBM (2007)	7
Data cleansing (data cleansing tool)	Cheong & Chang (2007)	2
Data cleansing	Todd (2008)	2
Data development	DAMA-DMBOK (2008)	2
Data development	Aiken et al.(2007)	2
Data integration	Lucas (2011)	2
Data integration	Aiken et al.(2007)	2
Data Management	Demchenko, De Laat, Membrey (2014)	1 4
Data management	Diché (2011)	1 4
Retention & archiving	Chapple (2013)	1 4
Document & Content Management	DAMA-DMBOK (2008)	1 4
Data taxonomy	Todd (2008)	1 4
Data traceability	Todd (2008)	1 4
Data requirements	Dyché (2011)	1 4
Data content	Hay (2014)	1 4
Data administration	Dyché (2011)	1 4
Data archiving	Todd (2008)	1 4

Data migration	Todd (2008)	1 4
Audit Information Logging & Reporting	IBM (2014)	1 4
Information management & usage	IBM (2007)	1 4
Third Party Data extract	Cheong & Chang (2007)	1 4
Data policies	Todd (2008)	4
Data policy	Lucas (2011)	4
Data principles	Khatri & Brown (2010)	4
Data standards	Todd (2008)	4
Data profiling	Cheong & Chang (2007)	3
Data profiling	Todd (2008)	3
Data profiling tool	Cheong & Chang (2007)	3
Data quality	Khatri & Brown (2010)	9
Data quality	Diché (2011)	9
Data quality communication strategies	Lucas (2011)	9
Data quality dimensions	Lucas (2011)	9
Data Quality Management	DAMA-DMBOK (2008)	9
Data Quality Management	IBM (2014)	9
Data quality management	Lucas (2011)	9
Data quality methodologies, technologies & tools	Lucas (2011)	9
Quality & consistency	Chapple (2013)	9
Data Risk Management	IBM (2014)	2
Data security management	DAMA-DMBOK (2008)	2
Governance metrics	Griffin (2010)	4
Metrics	Griffin (2011)	4
Metrics development & monitoring	Cheong & Chang (2007)	4
Data monitoring	Todd (2008)	4
Information Life-cycle management	IBM (2014)	5
Information lifecyle management	IBM (2007)	5
Data retention	Todd (2008)	5
Data retirement	Todd (2008)	5
Data lifecycle	Khatri & Brown (2010)	5
Master data management	DAMA-DMBOK (2008)	8
Master data management	Todd (2008)	8
Reference data management	Todd (2008)	8
Enterprise Data Model	IBM (2007)	8
Enterprise Data Stores	IBM (2007)	8
Data Warehousing	DAMA-DMBOK (2008)	8
Data model/types	Demchenko, De Laat, Membrey (2014)	8

Data modeling	Todd (2008)	8
Meta data management	DAMA-DMBOK (2008)	9
Metadata	IBM (2014)	9
Metadata	Khatri & Brown (2010)	9
Metadata Management	Diché (2011)	9
Metadata management	Cheong & Chang (2007)	9
Metadata management	Todd (2008)	9
Metadata repository	Cheong & Chang (2007)	9
Business metadata	Hay (2014)	9
Definitions of business metadata	Hay (2014)	9
Organisational bodies & policies	Cheong & Chang (2007)	2 3
Organization& policies	Griffin (2010)	2 3
Organizational structures (& Culture)	IBM (2007)	2 3
Organizational structures (& awareness)	IBM (2014)	2 3
People	De Abreu Faria, Maçada, Kumar (2013)	2 3
Policies & standards	Chapple (2013)	2 3
Policies (& practices)	De Abreu Faria, Maçada, Kumar (2013)	2 3
Policy	IBM (2014)	2 3
Principles & standards	Griffin (2010)	2 3
Processes& practices	Griffin (2010)	2 3
Sponsorship	Lucas (2011)	2 3
Governance policies	Diché (2011)	2 3
Governance Structure	Cheong & Chang (2007)	2 3
Roles, responsibilities & requirements	Griffin (2011)	2 3
Roadmap	Griffin (2011)	2 3
Executive Sponsorship	Diché (2011)	2 3
Decision rights	Cheong & Chang (2007)	2 3
Data program coordination	Aiken et al.(2007)	2 3
Data governance structure	Cheong & Chang (2007)	2 3
Data asset use	Aiken et al.(2007)	2 3
Business Model	Mohanty, Jagadeesh, Srivatsa (2013)	2 3
Issue escalation process	Cheong & Chang (2007)	2 3
User group charter	Cheong & Chang (2007)	2 3
Data Stewardship	Aiken et al.(2007)	2 3
Data stewardship	Diché (2011)	2 3
Data stewardship	Todd (2008)	2 3
Stewardship	IBM (2014)	2

		3
Data ownership	Todd (2008)	2 3
Data custodianship	Cheong & Chang (2007)	2 3
Security	Demchenko, De Laat, Membrey (2014)	5
Security	IBM (2014)	5
Security & Access rights	Diché (2011)	5
Security & Privacy	Chapple (2013)	5
Data privacy	Todd (2008)	5
Technology	IBM (2007)	7
Technology	Cheong & Chang (2007)	7
Technology	De Abreu Faria, Maçada, Kumar (2013)	7
Technology	Griffin (2010)	7
Technology	Chapple (2013)	7
Data storage	Todd (2008)	7
Infrastructure	Demchenko, De Laat, Membrey (2014)	7
Analytics	Demchenko, De Laat, Membrey (2014)	3
Benefits management & reporting	De Haes & Van Grembergen (2005)	3
Value Creation	IBM (2014)	3

Appendix D

Teradata New Regulations Outlined in "Principles for Effective Risk Data Aggregation and Risk Reporting" and Derived Platform Requirements (Kindler, 2013, pp.5)

NEW BASEL III GUIDELINES	DERIVED TECHNICAL PLATFORM REQUIREMENTS
Overarching governance and infrastructure	
A bank's board and senior management should promote the identification, assessment, and management of data-quality risks as part of its overall risk-management framework. The framework should include agreed service-level standards for both outsourced and in-house risk-data-related processes and a firm's policies on data confidentiality, integrity, and availability as well as risk-management policies.	• Clearly defined, implemented, and live data-governance policy • Clearly defined and guaranteed service levels for data processing, analysis, and reporting
A bank's risk data aggregation capabilities and risk-reporting practices should be fully documented and subject to high standards of validation. This validation should be independent and include review of compliance with the principles in this document. The validation should review the appropriateness and effectiveness of the bank's risk data aggregation capabilities and risk-reporting practices and the quality of the governance surrounding the processes. Independent validation could mean a review by the internal audit function. However, best practice would suggest that an independent validation unit with specific IT, data, and reporting knowledge may be better positioned to perform this review.	• Review of architecture, effectiveness, and compliance by an external and independent validation unit with specific IT, data, and reporting knowledge
A bank should design, build, and maintain data architecture and IT infrastructure that fully supports its risk data aggregation capabilities and risk-reporting practices, not only in normal times but also during times of stress or crisis, while still meeting the other principles.	• Risk-architecture, analysis, and reporting capabilities outlined and scaled for worst-case conditions • Infrastructure scaled to max but payment for utilization only
Risk data aggregation capabilities	
A bank should be able to generate accurate and reliable risk data to meet normal and stress/crisis reporting requirements. Data should be aggregated on a largely automated basis so as to minimize the probability of errors. Risk data should be reconciled to accounting data as well as to a bank's sources and books of record to ensure that the risk data is accurate.	• Accuracy of reporting under stress/crisis • Automated data sourcing and aggregation, minimal manual interaction • Reconciled finance and risk data • Common data model for finance and risk • Ideally, shared data warehouse for finance and risk
A bank should strive toward a single authoritative source for risk data.	• One source of data for risk data aggregation and reporting • One source of truth

NEW BASEL III GUIDELINES	DERIVED TECHNICAL PLATFORM REQUIREMENTS
As a precondition, a bank should have a dictionary of the concepts used, such that data is defined consistently across an organization.	• One logical data model across the risk and finance area • One business data model (access layer, etc.) across the risk and finance area
Supervisors expect banks to develop metrics to monitor the accuracy of data and to have appropriate escalation channels and action plans in place to rectify poor data quality. Supervisors could expect banks to monitor and report on the number of data items that are received, compared to the number of items expected.	• High data quality • Data-quality metrics • Automated data-quality monitoring
A bank should be able to capture and aggregate all material risk data across the banking group. Data should be available by business line, legal entity, asset type, industry, region, and other groupings that permit identifying and reporting risk exposures, concentrations, and emerging risks.	• Central data warehouse with all data from all divisions within the bank • Data storing in lowest granularity level to enable aggregation across different dimensions
A bank should be able to generate aggregate and up-to-date risk data in a timely manner while also meeting the principles relating to accuracy and integrity, completeness, and adaptability. The precise timing will depend upon the nature and potential volatility of the risk being measured as well as its criticality to the overall risk profile of the bank. This timeliness should meet bank-established frequency requirements for normal and stress/crisis risk-management reporting. Critical risks include but are not limited to operational risk indicators that are time critical (e.g., systems availability, unauthorized access).	• Timely import of new data to data warehouse • Rapid production of new analysis and reports (depending on criticality of results) • Intraday data on-demand import, aggregation, analysis, and reporting • System-log analysis resulting in required unstructured data-analysis tools and big-data requirements
A bank should be able to generate aggregate risk data to meet a broad range of on-demand, ad-hoc risk-management reporting requests, including requests during crisis situations, requests due to changing internal needs, and requests to meet supervisory queries. A bank's risk data aggregation capabilities should be flexible and forward-looking to assess emerging risks. Adaptability will enable banks to conduct better risk management, including forecasting information as well as supporting stress testing and scenario analyses. Adaptability includes capabilities for data customization to users' needs (e.g., dashboards, key takeaways, anomalies), to drill down as needed, and to produce "flash" summary reports.	• Flexibility in implementation of new requirements • Rapid time to market for new requirements • Ad-hoc queries • Ad-hoc analysis besides standard reporting • Forward-looking analysis and scenario calculations • Ad-hoc prediction of future risk • Interactive stress testing across all data and risk factors of the bank across all dimensions • Drill-down capabilities to lowest granularity • Rapid visualization of ad-hoc results
A bank should be able to incorporate new developments on the organization of the business or external factors that influence the bank's risk profile or the requirements to measure its components and capabilities to incorporate changes in the regulatory framework.	• Rapid business-driven change and enhancement capabilities within entire risk-aggregation value chain
Supervisors expect banks to be able to generate subsets of data based on requested scenarios or resulting from economic events. For example, a bank should be able to aggregate risk data quickly on country credit exposures as of a specified date based on a list of countries, as well as industry credit exposures as of a specified date based on a list of industry types across all business lines and geographic areas.	• Ad-hoc scenario analysis capabilities on any set of data across the entire bank • Full business-driven flexibility in setting up new simulations • Drill-down capabilities on ad-hoc scenario simulations

NEW BASEL III GUIDELINES	DERIVED TECHNICAL PLATFORM REQUIREMENTS
Risk-reporting practices	
Accurate, complete, and timely data is a foundation for effective risk management. However, data alone does not guarantee that the board and senior management will receive appropriate information to make effective decisions about risk. To manage risk effectively, the right information needs to be presented to the right people at the right time. Risk reports based on risk data should be accurate, clear, and complete. They should contain the correct content and be presented to the appropriate decision makers in a time frame that allows for an appropriate response. A bank's risk-management reports should contribute to sound risk management and decision making by their relevant recipients, including, in particular, the board and senior management.	• Flexibility to send the right information at the right time to the right people
Risk-management reports should accurately and precisely convey aggregated risk data and reflect risk in an exact manner. Reports should be reconciled and validated.	• Reconciliation capabilities across different results
To ensure the accuracy of the reports, a bank should maintain automated and manual edit and reasonableness checks, including an inventory of the validation rules that are applied to quantitative information. The inventory should include explanations of the conventions used to describe any mathematical or logical relationships that should be verified through these validations or checks.	• Automated reasonability and quality checks
Risk-management reports should cover all material risk areas within the organization. The depth and scope of these reports should be consistent with the size and complexity of the bank's operations and risk profile, as well as the requirements of the recipients. Risk-management reports should include exposure and position information for all significant risk areas (e.g., credit, market, liquidity, operational) and all significant components of those risk areas (e.g., single name, country, and industry sector for credit risk). Risk-management reports should also cover risk-related measures (e.g., regulatory and economic capital).	• All material risk data within the organization included in data aggregation and analysis • All transactional data produced within a bank included in the risk data warehouse
Supervisors expect that risk-management reports will be complete. A bank should determine risk-reporting requirements to best suit its own business models and risk profiles. Supervisors will need to be satisfied with the choices a bank makes in terms of risk coverage, analysis and interpretation, scalability, and comparability across group institutions. For example, an aggregated risk report should include, but not be limited to, capital adequacy, regulatory capital, capital and liquidity ratio projections, credit risk, market risk, operational risk, liquidity risk, stress testing results, inter- and intra-risk concentrations, and funding positions and plans	• Risk results comparable across the entire organization and all divisions • One data model (logical, semantic layer, data access layer, etc.) across the organization and all divisions
Supervisors expect that risk-management reports to the board and senior management provide a forward-looking assessment of risk and should not just rely on current and past data. The reports should contain forecasts or scenarios for key market variables and the effects on the bank so as to inform the board and senior management of the likely trajectory of the bank's capital and risk profile in the future.	• Ex-post-analysis and an ex-ante simulation layer available across all risks within the bank
The board and senior management (or other recipients as appropriate) should set the frequency of risk-management report production and distribution. Frequency requirements should reflect the needs of the recipients, the nature of the risk reported, and the speed at which the risk can change as well as the importance of reports in contributing to sound risk management and effective/efficient decision making across the bank. The frequency of reports should be increased during times of crisis.	• Analysis and reporting frequency to match the speed with which the underlying risk may change • Capability for the reality that credit risk, market risk, and liquidity risk all depend on capital market prices and can change drastically within seconds • Intra-day risk reporting at a minimum or, better, within hours or minutes • Risk on demand in times of market turmoil

NEW BASEL III GUIDELINES	DERIVED TECHNICAL PLATFORM REQUIREMENTS
Supervisory review, tools, and cooperation	
Supervisors should periodically review and evaluate a bank's compliance with the Basel III principles. Reviews should be incorporated into the regular program of supervisory reviews and may be supplemented by thematic reviews covering multiple banks, with respect to a single or selected issue. Supervisors may test a bank's compliance with the principles through occasional requests for information to be provided on selected risk issues (e.g., exposures to certain risk factors) within short deadlines, thereby testing the capacity of a bank to aggregate risk data rapidly and produce risk reports. Supervisors should have access to the appropriate reports to be able to perform this review.	• Capability for more frequent and rapid review and testing of aggregation and analytical results by regulators • Ability to explain data and analytical results produced in the past • Rapid retrievability of historic data content and assumptions going into analysis • Temporal database design and retrieval capabilities
Supervisors should draw on reviews conducted by the internal and external auditors to inform their assessments of compliance with the principles. Supervisors may require work to be carried out by a bank's internal audit functions or by experts independent from the bank. Supervisors must have access to all appropriate documents such as internal validation and audit reports and should be able to meet with and discuss risk data aggregation capabilities with the external auditors when appropriate.	• Review and assessments of data aggregation and analysis from external experts

Appendix E

Ranking of data governance elements based on the Basel III framework

Process	Sub-process	Frequency in Basel III principles & guidelines
Roles, structures & policies	• Culture and awareness	0
	• People	0
	• Policies and standards	1
	• Business model	0
	• Processes & practices	1
	• Data stewardship	0
Data management	• Document and content management	2
	• Retention and archiving management	2
	• Data traceability	2
	• Data taxonomy	2
	• Data migration	2
	• Third party data extract	2
	• Data storage	2
Data quality management	• Quality methodologies and tools definition	2
	• Quality dimensions	2
	• Quality communication strategies	2
Metadata management	• Definitions of business metadata	1
	• Metadata repository	1
Master data management	• Reference data management	1
	• Data modeling	2
	• Enterprise data model	2
	• Data stores	1
	• Data warehousing	2
	• Data integration	3
Data architecture	• Data entity/data component catalog	1
	• Data entity/business function matrix	1
	• Application/data matrix	1
	• Data architecture definition	1
Technology	• Infrastructure	2
	• Analytics	5
	• Business applications	1
Security & privacy	• Data access rights	1
	• Data risk management	1
	• Data compliance	2
Metrics development and monitoring	• Benefits management & monitoring	1
	• Value creation quantification	1