CHAPTER 5

BIG DATA GOVERNANCE

Malgorzata PANKOWSKA*

*Professor, University of Economics in Katowice, Faculty of Informatics & Communication, Department of Informatics, Katowice, Poland

E-mail: pank@ue.katowice.pl

DOI: 10.26650/B/ET06.2020.011.05

Abstract

Information processing in a traditional way focuses on relatively stable structured data, repeatable processes as well as on operations in Business Intelligence systems. However, nowadays more and more popular, big data, defined as huge volumes of data available in varying degrees of complexity, generated at different velocities, and varying degrees of ambiguity, cannot be processed using traditional methods and technologies. Some people argue that suitable IT (Information Technology) infrastructure for big data processing is not yet widely developed nor implemented to discuss the big data architecture implementation benefits, risks, and opportunities. Nevertheless, this paper is to present the big data governance issues. Particularly, within the proposed theme, the author discusses the big data system architecture and development strategy. The last part of the paper includes a proposal of a big data architecture model as well as a design of balanced scorecard objectives and measures specification to support the big data governance at public services business organizations. As usual, there are two main research methods, i.e., literature review and the analysis of case studies. The first provides an overview of the existing knowledge and the second permits for contextualization of the proposed models. Beyond that, the paper includes definitions of the key concepts and enables to extend the knowledge base in the research area.

Keywords: Big data, System architecture, ArchiMate, Data governance, Balanced scorecard

1. Introduction to Big Data Issues

The term of big data was coined to describe large data repositories because of the explosive increase of the amount of global data. Its tremendous growth has come from people's daily life, especially in relation to the Internet, social media and mobile devices. In general, big data concerns the datasets, which could not be perceived, acquired, managed, and processed by traditional software tools within an acceptable time and scope. Big data includes data from blogs, tweets, social networking sites, news feeds, discussion boards, video sites, web logs in various, semi-structured formats, as well as machine-generated data, from RFIDs and other sensors, such as optical, acoustic, thermal, seismic, chemical and medical devices. Other examples include shoppers' smartphones, road sensors, GPS devices, TV set boxes, and video cameras. Gathering big data requires modifying the traditional view of the data warehouses. Business organizations are now involved in mixing structured, unstructured and streaming data that often has low latency requirements and still supports queries. Therefore, they need new IT for processing. Business organizations recognize that there is a wealth of data in open social media. Nonetheless, social media data is unstructured and requires different technologies to process and extract useful information.

The objective of data governing for a business organization is to process data as quickly as it is technologically possible while keeping the quality as high as it is practically possible. Traditional approach to data processing concerns processing of transactions and the Internet application data as well as mainframe, OLTP (online transaction processing) system or ERP (enterprise resource planning) system data. For years, data has been stored in transaction and database systems, and in data warehouses. The new approach to data processing is to be creative, holistic, and intuitive, because data is from different sources. According to Krishnan (2013) the basic premises of big data architecture modelling and implementation cover business model transformation by globalization and connectivity, personalization of services, communication media convergence and new sources of data because of the advances in mobile technology, large-scale data processing networks, commoditization of hardware, virtualization, cloud and fog computing. Considering that big data is neither structured, nor does it have a finite state and volume, the complexity of big data relies on the following (Dong & Srivastava, 2015, Gupta & Singla, 2017):

- Data volume, which reflects the amount of unique data converting low-density data into high-density data that has value. The amount of data varies from organization to organization ranging from terabytes to petabytes;

- Data velocity, which is the rate for receiving data. The highest velocity data streams directly into computer memory instead of being written to disk. Many mobile Internet applications enable real-time evaluation and action.
- Data variability that is dissimilar from variety. Almost the same message is transferred in different communication channels, eventually to the same person but in a slightly different form. The meaning of the message constantly changes if the information context changes. Furthermorethe information recipients change their locations or other specific characteristics.
- Data variety, which means that data comes in all types of formats from structured, numeric data in traditional databases to unstructured text documents, emails, video, audio and financial transactions.
- Data value, which is expected to be calculated. Value is derived by a range of quantitative and investigative techniques. The cost of data storage and computation has rapidly decreased because of IT development, but finding the data value requires creation and implementation of analytical processes and involves business analysts, users, and executives. They are learning to ask the right questions, recognizing patterns, and predicting phenomena or behaviours.
- Data veracity that concerns the exactness and precision of data as well as the dependability of information. Because of many types of enormous data, quality and exactness are less controllable, however, the big data investigation innovation allows to work with these different kinds of data as reliable.
- Data visualization, which is critical for data usage. Using charts and graphs to visualize large amount of data is nowadays much more effective than some years ago.

Chen et al. (2014) emphasize obstacles in the development of big data applications:

- Data representation. Many datasets have certain levels of heterogeneity in type, structure, semantics, granularity, organization, provenance, and accessibility, therefore data stewards and managers need techniques and tools to integrate the data.
- High degree of redundancy in datasets, therefore redundancy reduction and data compression tools are expected to be developed.
- Data analytics models, techniques, and tools are constantly required to be developed and implemented because of data volume and velocity.

Big data analytics can be defined as a combination of traditional analytics and data mining techniques along with huge volumes of data to create a platform to analyze, model, and predict the behaviour of customers, markets, products, services, and the competition to enable the achievement of competitive advantage on the market. Traditional data analysis means to use appropriate statistical methods to analyze massive data to concentrate, extract, and refine the useful data from the chaotic. Many traditional data analysis methods are still valid in the new big data environment, e.g., cluster analysis, factor analysis, correlation analysis, regression analysis, but also new techniques of data mining and decision support are implemented. According to Schmarzo (2013), the biggest difference between the business intelligence (BI) analyst and data scientist is the environment, in which they work. BI professional is working in a highly structured data warehouse environment. This environment is typically product or market driven, with highly centralized management of IT services and service level agreements (SLA) implementation in order to ensure timely generation of managerial dashboards and reports. On the other hand, data scientists create separate data sandboxes to load whatever data they can get on both internal and external data sources, and later on they are involved in data cleansing, profiling, transformation, creation of new metrics and models, and testing. Morabito et al. (2015) formulated the process of big data analytics, which included six steps and required specific policies and procedures due to the characteristics of big data:

- Identification of key priorities and recognition of the business context, and setting up the analytics goals.
- Selection of the appropriate data for the analysis.
- Enhancement of data reliability by defining missing data, or removal of irrelevant data and outliers, as well as compilation of data coming from different sources.
- Data mining to verify hypotheses and to extract the meaningful signals.
- Evaluation of data processing results and pattern interpretations.
- Visualisation and reporting on the achieved results.

Data science takes advantage of big data because of its exceptional scale and possibilities to process heterogeneous data, i.e., texts, images, graphs, or sounds. Data analytics allow for deeper insights into the data and improving the quality of products and services offered by business organizations through its multidisciplinary functionality. There is a need to emphasize the difference between big data and data science. Big data is a term

used to concern the exponential growth and availability of data, which can be structured or unstructured. Data science is a research field on knowledge drawn from large volumes of heterogeneous data (i.e., video, audio, text and image). Data science is connected with data analysis, statistics, machine learning and data mining as well as knowledge discovery in databases. Although business analytics cover the use of data-driven insight to generate value, the big data architecture governance is necessary to enable business analytics as well as data science research.

2. Big Data Architecture

Big data analytics is accepted as a very attractive research domain with significant impact on industrial and scientific domains. Belcastro et al. (2017) identified the key research sub-fields, which cover programming models for big data analytics, data storage scalability, data availability by cloud service providers, data interoperability and openness, data quality and usability improvement, integration of big data analytics frameworks, development of tools for massive social network analysis, local mining and distributed model coordination, and in-memory analysis. That research challenges are required to be supported by appropriate system architecture. According to Azarmi (2016) the system architecture modelling should take into account some common issues to create the right sizing. The important ones comprise of defining the appropriate size of daily data input, the structure of the ingested data, the average number of events ingested per second, the retention period, the required availability of data, the expected indexing throughput, the number of visualization users, or the centralization of logs management. System architecture developers, bearing these requirements, already proposed some referential models of system architecture. Therefore, the NIST model can be considered as fundamental for big data architecture. As Heisterberg and Verma (2014) argue that people determine business architecture, process-application architecture, and tools-technology architecture, the NIST model defines three cloud service models appropriate for big data:

- Infrastructure as a Service (IaaS). This includes the storage, servers, and network as the base. The distributed file systems are part of this layer, therefore the big data is also stored in cloud repositories.
- Platform as a Service (PaaS). The NoSQL data storages and distributed caches that can be logically queried using query languages form the platform layer of big data. The layer includes NoSQL and relational databases.

- Software as a Service (SaaS). Specific industries, like health, retail, e-commerce, energy, or banking can build packaged applications that serve a specific business need and leverage the data for cross-cutting data functions.

Heisterberg and Verma (2014) argue that big data architecture is expected to address all type of data coming from various data sources, such as enterprise applications. There is data generated from ERP (enterprise resource planning) systems, CRM (customer relationship management) systems, SCM (supply chain management) systems, e-commerce transactions, HR (human resources) and payroll transactions. Beyond that, there are records from call centers, web logs, smart meters and manufacturing sensors data, equipment logs, and trading systems data generated by machine and computer systems. Companies oriented towards social networking collect social media data, which covers customer feedback streams, microblogging sites like Twitter, and social media platforms like Facebook data.

Big data system architecture is to process business vision and strategy into effective enterprise by creating, communicating and improving the key principles and models that describe the enterprise's future state and enable its continuous transformation. Unhelkar (2018) emphasized some essential advantages of big data architecture development, such as creating a positive impact on the agility of business, expanding new horizons for data analytics and technologies, collaborations by shareable architecture and by involvement in cloud computing. Beyond that, sustainable computing and environmental considerations in business operations are also opportunities made possible through big data architecture. Figure 1 covers big data system architecture model for public services business organization. The proposed model of architecture for realizing big data solutions includes heterogeneous infrastructures, databases, data repositories, and visualization and analytics tools. Many open source frameworks, databases, Hadoop distributions, and analytics tools are available on the market, however, introducing big data requires firstly to answer the question of why it is necessary. The answer is included in the Motivation layer in the model architecture in Figure 1. In general, the proposed model consists of four layers in ArchiMate language and in OMG free software tool. Everything starts from the top layer, which is the Motivation layer, covering the identification of business stakeholders, goals, drivers, principles, and assessments. The model of architecture is to be always allocated in certain context. In this case study, the public service business organization context is proposed. The second layer in Figure 2 is the Business layer covering business processes of public service organization as well as big data management processes. Data management consists of two primary groups of activities, i.e., the management of organization-wide conceptual data models and the

management of organization-wide data standards. Data manager is deputed to be responsible for organization-wide coordination, focusing on the goals and plans for data quality management among responsible organization units. The third Application layer in Figure 1 compromises all applications for data processing. The traditional approach to data management is based on centralized assembling all the company data. However, big data is retained in distributed systems instead of a centralized one. As it is presented in big data architecture model, data is stored in databases, data marts, data warehouses as well as in data lakes. The last term refers to the container for raw data. Data lake is a system that stores data from a single source. However, today data lake is considered a general, enterprise-wide data repository for data from multiple sources (Quix and Hai, 2019).

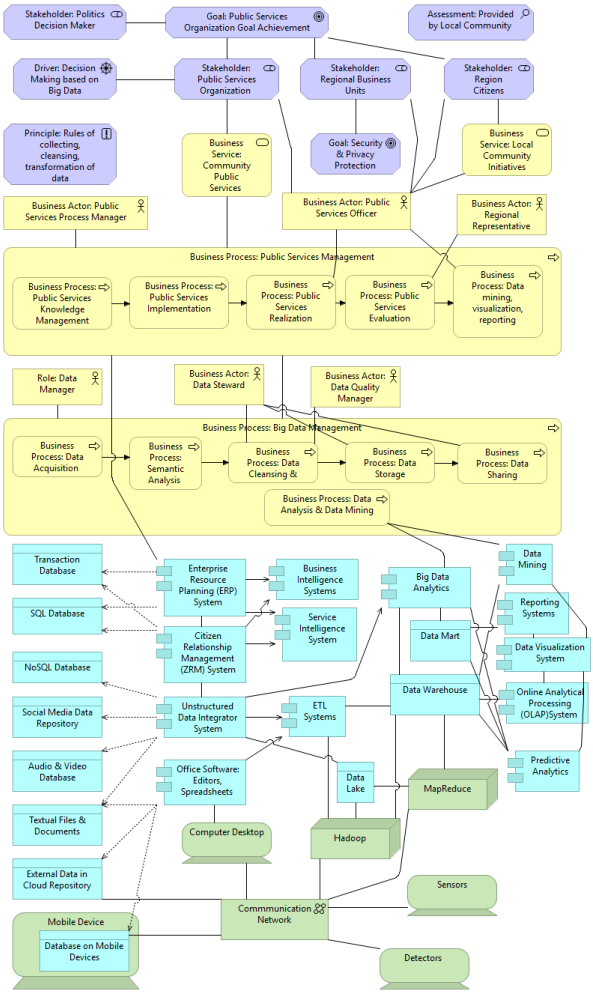

Figure 1: Big data architecture model for public service business organization

3. Big Data Balanced Scorecard

Big data governance leaders provide a framework for setting data-usage policies and controlling if information remains accurate, consistent and accessible. The strong point in their activity is to design the big data architecture model and implement it. However, this model is usually supplemented by additional elaborations, i.e., business strategy and IT strategy. Both of them are supported by a balanced scorecard (BSC), a tool proposed by Kaplan and Norton (2004).

In this paper, big data governance is understood as a developed and enforced set of processes to ensure that important data assets are formally and consistently managed across different heterogeneous platforms to ensure the required level of data quality. Melgarejo Galvan and Navarro (2017) identified problems typical for data governance. These problems particularly refer to the speed of processing, confidentiality of data, ignoring public sources, which contain unstructured, but valuable data . The other problem concerns unstandardised and inconsistent data, which should be cleaned up before the process of analysis. In general, big data governance is oriented towards the increase of processed data value. The process requires that users know what questions they want to answer and for what decisions. This knowledge permits to uncover new opportunities that impact business activities, reduce costs, and mitigate risks across all operational and financial aspects of an organization's value chain.

According to Plotkin (2014), data governance is the authority and the exercise of decision making. Simultaneously, it is considered as a system of decision rights and accountabilities for processes, executed according to agreed models. DAMA International Guide to the Data Management Body of Knowledge (Cupoli et al., 2014) defines the data governance main domains as follows: data architecture, data quality, data modeling and design, data storage and operations, data security, data warehousing and business intelligence, data integration and interoperability, and meta-data management. Also, Smallwood (2014) emphasized big data governance as a key concern in the environment of big data because of data risk management, privacy and security, business operations long-term digital preservation, and business intelligence including sub disciplines like document management, knowledge management, business continuity and disaster recovery. Big data governance can be considered as an organizational challenge, i.e., roles have to be identified, stakeholders have to be assigned to roles and responsibilities, and business processes need to be established to organize various issues around data governance. Some of these issues are already solved at the stage of big data architecture modeling and implementation. Big data governance should

be supported by appropriate techniques and tools. Beyond the above mentioned subtopics, big data governance requires defining data ownership, stewardship and protection. Anilkumar et al. (2017) emphasize meta-data and master data management, data dictionaries and standards maintenance, audit validation, and data life cycle management. Van Helvoirt and Weigand (2015) argue that master data management (MDM) adds a new value to the data, because MDM focuses on establishing integration and interoperability of heterogeneous databases and applications in a business oriented manner. Following Van Helvoirt and Weigand (2015), big data governance is presumed to be more than just achieving compliance, because it is necessary to adopt practices and principles that increase data quality and trust. A valuable data quality standard series is ISO 8000, which focuses on data characteristics and exchange in terms of vocabulary, syntax, semantics, encoding, provenance, accuracy and completeness. Standard ISO/TS 8000-1:2011 contains an introduction to ISO 8000. It covers a statement of the scope of ISO 8000, principles of data quality, the high-level data architecture of ISO 8000, a description of the ISO 8000 structure, and a summary of the content of the other parts of the general data quality series of parts of ISO 8000. Standard ISO 8000-2:2018 Data quality –Part 2: Vocabulary enables to create, collect, store, maintain, transfer, process and present data to support business processes effectively. Standard ISO 8000:150 includes a framework for data quality management. This model comprises processes, named data operations, data quality monitoring, and data quality improvement, which are in general oriented towards constant improvement of quality management. These processes are connected with particular roles, i.e., data manager, data administrator and data technician, who are responsible for process activities. According to Unhelkar (2018) quality practices in big data domain cover data profiling, cleansing and standardizing the data, applying syntax, semantics, and aesthetic checks to data, using standard architectural reference models and data patterns, controlling the business processes quality, continuous testing, and using agile techniques in developing high-quality analytics. Data cleansing consists of finding errors in data, removing unnecessary duplications, inconsistencies or incomplete values. The data is corrected by replacing values generated or deleted in the worst case (Melgarego Galvan and Navarro, 2017).

Although business analytics concerns the use of data-driven research to generate value, the big data architecture and big data governance are necessary to enable the business analytics as well as the data science research. In business organization the value is created through leveraging people, processes, data and technology. These components should be included in big data governance models. Encompassing all of these elements is the

organization culture, as a system of shared values. People are the professionals and their skills are involved in applying business analytics. Processes are a series of activities linked to achieve an outcome, like information required by decision maker. All these assets, i.e., people, processes, data and technology are applied to achieve value. As Chi (2015) notices, while collecting data is not difficult, value creation out of the data is still often questionable. Although the data intensive scientific discoveries are more and more published, and huge big data repositories are developed, business organization still have a question, if they really need the data for decision making. Therefore, this paper covers a proposal of a balanced scorecard (BSC) as a tool to support the decision making of strategic investment in big data architecture implementation. In the proposed balanced scorecard (Figure 2) four perspectives are included:

- Financial Perspective covering financial measures like return on investment (ROI), revenues, costs, and business model to increase adherence to audit and corporate responsibility, reduce time to market, and reduce complaints.
- Social Perspective concerning leadership support, professional involvement, strategic alliances and partnerships for enhancing business analytics capabilities, training and continuous learning.
- Process Perspective including descriptive, prescriptive, and predictive analytics, used to understand and solve specific problems.
- Technology Perspective conveying information in reports and dashboards, and including cross-department applications, substantial IT infrastructure, data marts, data lakes, data warehouses, Hadoop and MapReduce technology, visualization tools, cloud storage, mobile applications, and advanced machine learning tools.

As well as corporate governance, the big data governance is the organizational capacity to control the formulation and implementation of big data strategy and in this way ensure the fusion of business and big data. Van Grembergen and DeHaes (2008) argue that IT management is to be included in the IT governance process. They emphasize the role of the governance process as value creation. They assume that focus area for big data governance is driven by stakeholder values like strategic business – information technology alignment (BITA), resource management, risk management and performance management.

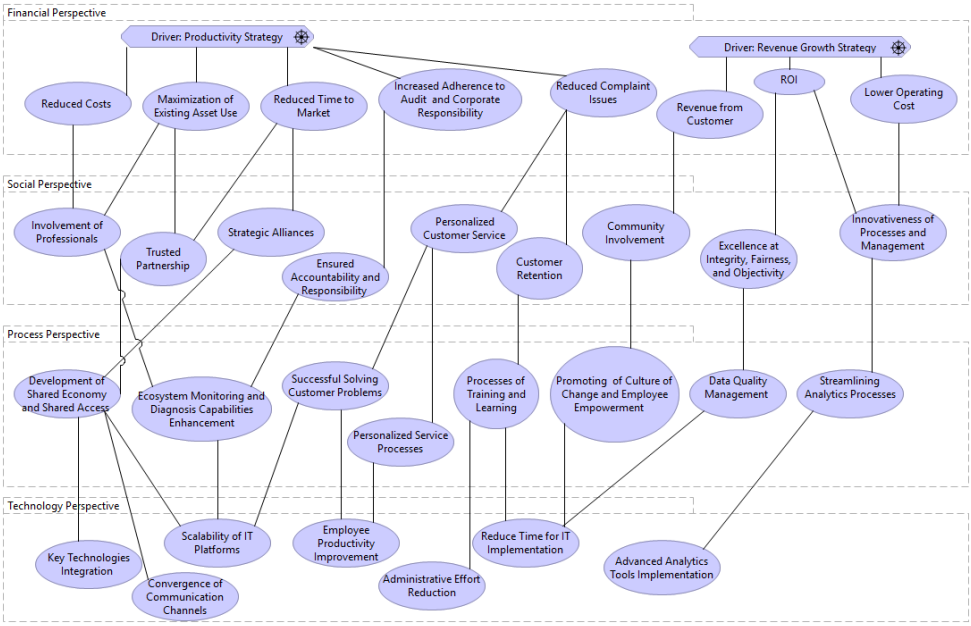

Figure 2: Big Data Balanced Scorecard Perspective Objectives

Balanced scorecard is defined to balance short term cost savings with long term performance, optimally match big data resources to business requirements, manage big data security and risk to business (Kaplan and Norton, 2004). Taking into account that BSC focuses on the value of assets and activities provided, as well as on eliminating non-value adding activities and processes, in Figure 2 the proposed values for big data governance are presented. The proposals are to emphasize what seems to be extremely important from each of the particular perspectives. The proposed valuable objectives can be linked among perspectives and in this way the compatibility of perspectives can be presented. Therefore, the balanced scorecard offers a framework for the development of big data implementation strategies for creating values. Proposed balanced scorecard is offered as a referential model, which is expected to encourage public services business organization to consider its decision on the big data architecture development, investment and implementation. If the proposed objectives can be achieved then the big data architecture should be successfully implemented.

4. Conclusion

Stimmel (2015) argues that although analyzing big data ensures advantages for many business organizations, providing effective governance is not an easy task. In the paper, big data governance was considered as a discipline, not a particular activity. This discipline is

value oriented and the balanced scorecard could be applied as a referential model to support that discipline decision making.

As proposed by Kaplan and Norton (2004) the BSC can be applied when a value creation is not direct. Intangible assets such as knowledge, big data, business analytics have an indirect impact on financial and social outcomes such as increased productivity, revenues, lowered costs and time to markets, and higher profits and customer satisfaction. Although values are contextual, referential models can be considered as helpful in that values' contextualization for each particular case.

References

Anilkumar, R., Deshmukh, R.R., Emmanuel, M. (2017) Big Data Predictive Analysis for Detection of Prostate Cancer on Cloud-Based Platform: Microsoft Azure, Privacy and Security Policies. In Tamane, S., Kumar Solanki, V., Dey, N. (eds.) *Privacy and Security Policies in Big Data*. IGI Global, Hershey, 259-278

Azarmi, B. (2016) *Scalable Big Data Architecture, A practitioner's guide to choosing relevant big data architecture*. Springer NY.

Belcastro, L., Marozzo, F., Talia, D., Trunfio, P. (2017) Big Data Analysis on Clouds. In Zomaya A.Y., Sakr S. (eds.) *Handbook of Big Data Technologies*. Springer Cham, 101-142.

Chen, M., Mao, S., Zhang, Y., Leung, V.M. (2014) *Big Data, Related technologies, challenges and future prospects*. Springer, Cham Heidelberg.

Chi, C-H.(2015) Behaviour Informatics: Capturing Value Creation in the Era of Big Data. In Intan, R., Chi, C-H., Palit, H.N., Santoso, L.W. (eds.) *Intelligence in the Era of Big Data*. Springer Verlag Berlin, XIV-XVI.

Cupoli, P., Earley, S., Henderson, D. (2014) *DAMA -DMBOK2 Framework, The Data Management Association*. https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf. Accessed July 12, 2019.

Dong, X.L., Srivastava, D. (2015) *Big Data Integration.* Morgan & Claypool Publishers, Waterloo.

Gupta, M., Singla, N. (2017) Evolution of Cloud in Big Data With Hadoop on Docker Platform. In Tamane, S., Kumar Solanki, V., Dey, N. (eds.) *Privacy and Security Policies in Big Data*. IGI Global, Hershey, 41-64.

Heisterberg, R., Verma, A.(2014) *Creating Business Agility, How Convergence of Cloud, Social, Mobile, Video, and Big Data Enables Competitive Advantage*. Wiley, Hoboken.

ISO/TS 8000-1:2011, Data quality –Part 1:Overview, https://www.iso.org/standard/50798.html. Accessed July 12, 2019.

ISO/TS 8000:150:2011, Data quality –Part 150: Master data: Quality management framework, https://www.iso.org/standard/54579.html. Accessed July 12, 2019.

ISO 8000-2:2018 Data quality –Part 2: Vocabulary, https://www.iso.org/standard/76563.html. Accessed July 12, 2019.

Kaplan, R.S., Norton, D.P. (2004) *Strategy Maps, converting intangible assets into tangible outcomes*. Harvard Business School Press, Boston.

Krishnan, K.(2013) *Data Warehousing in the age of Big Data*. Morgan Kaufmann, Elsevier, Amsterdam.

Melgarejo Galvan, A.R., Rocio Clavo Navarro, K. (2017) Big Data Architecture for Predicting Churn Risk in Mobile Phone Companies. In Lossio-Ventura, J.A., Alatrista-Salas, H. (eds.) *Information Management and Big Data*. Springer Heidelberg, 120-133.

Morabito, V. (2015) *Big Data and Analytics, Strategic and Organizational Impacts.* Springer Cham.

Plotkin, D. (2014) *Data Stewardship, An Actionable Guide to Effective Data Management and Data Governance*. Elsevier, Amsterdam.

Quix, Ch., Hai, R. (2019) *Data Lake*. In Sakr, S., Zomaya, A.Y (eds.) *Encyclopedia of Big Data Technologies*. Springer Nature, Cham, 552-559.

Schmarzo, B. (2013) *Big Data, Understanding How Data Powers Big Business.* Wiley, Indianapolis.

Smallwood, R.F. (2014) *Information Governance*. John Wiley and Sons, Hoboken.

Stimmel, C.L. (2015) *Big Data Analytics Strategies for the Smart Grid*. CRC Press, Taylor & Francis Group, London.

Unhelkar, B. (2018) *Big Data Strategies for Agile Business, Framework, Practices and Transformation Roadmap*. CRC Press, Boca Raton, London.

Van Grembergen, W., DeHaes, S. (2008) *Implementing Information Technology Governance, Models, Practices, and Cases*. IGI Publishing, Hershey, New York.

van Helvoirt, S., Weigand, H. (2015) Operationalizing Data Governance via Multi-level Metadata Management. In Janssen, M., Mäntymäki, M., Hidders, J., Klievink, B., Lamersdorf, W., van Loenen, B., Zuiderwijk, A. (eds.) *Open and Big Data Management and Innovation*. Springer, Cham, 160-172.