Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects: 13th International Conference, KMO 2018, Žilina, Slovakia,...

Chapter · July 2018

DOI: 10.1007/978-3-319-95204-8_51

CITATIONS
25

READS
9,098

5 authors, including:

Silvia Lozano
Universidad EAFIT
3 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Juan David Ospina Arango
National University of Colombia
82 PUBLICATIONS **1,013** CITATIONS
SEE PROFILE

Marta Tabares
Universidad EAFIT
37 PUBLICATIONS **113** CITATIONS
SEE PROFILE

Towards an improved ASUM-DM process methodology for cross-disciplinary multi-organization Big Data & Analytics projects

Santiago Angée[1], Silvia I. Lozano-Argel[1], Edwin N. Montoya-Munera[1], Juan-David Ospina-Arango[2], Marta S. Tabares-Betancur[1]

[1] Eafit University, Medellin, Colombia
[2] Bancolombia S.A., Medellin, Colombia
{sangeea, slozanoa, emontoya, mtabares}@eafit.edu.co
judaospi@bancolombia.com.co

Abstract. The development of big data & analytics projects with the participation of several corporate divisions and research groups within and among organizations is a non-trivial problem and requires well-defined roles and processes. Since there is no accepted standard for the implementation of big data & analytics projects, project managers have to either adapt an existing data mining process methodology or create a new one. This work presents a use case for a big data & analytics project for the banking sector. The authors found out that an adaptation of ASUM-DM, a refined CRISP-DM, with the addition of big data analysis, application prototyping, and prototype evaluation, plus a strong project management work with an emphasis in communications proved the best solution to develop a cross-disciplinary, multi-organization, geographically-distributed big data & analytics project.

Keywords: big data, analytics, project management, CRISP-DM, ASUM-DM, process methodology

1 Introduction

The development of a big data & analytics project with the participation of several corporate divisions within an organization is a non-trivial problem and requires well defined roles and processes. The complexity of this kind of projects rises mainly when more than one organization participates in the project and these organizations are geographically distributed [13]. Furthermore, since there is no accepted standard for the implementation of big data & analytics projects so far, some authors point out that the project managers have to either adapt to existing data mining/predictive analytics methodologies, such as the *CRoss Industry Standard Process for Data Mining* (CRISP-DM) created by IBM [4] or *Sample, Explore, Modify, Model, and Assess* (SEMMA) created by SAS [12], or create a new one [17, 20].

Nonetheless, the usage of CRISP-DM and SEMMA is decreasing [19] and this is, in part, because these process methodologies focus specifically on the improvement of the technical components of the process and not on other project's dimensions such as knowledge management, communication management and project management, which are also crucial for a big data & analytics project's success [1]. Yet, a poll conducted by Piatetsky showed that CRISP-DM is still the most used methodology for data mining and predictive analytics projects [17]. Hence, the authors found out that there is an increasing necessity to adapt an existing process methodology, so that the aforementioned dimensions plus the technical components of the project can be integrated to increase the probability of success of a big data & analytics project.

This work presents a use case of a big data project that made use of an adaptation of the *Analytics Solutions Unified Method for Data Mining/predictive analytics* (ASUM-DM) process methodology, a refined CRISP-DM [9]. This use case shows how the ASUM-DM-adapted process methodology helped address the project's big data peculiarities, including the project management, communication management, technical and knowledge management issues.

This study comprises the first phases of the ASUM-DM process methodology: Analyze, Design, Configure and Build, along with Project management, and illustrates how an adaptation of ASUM-DM can help solve the project management, communication management, technical and knowledge management issues that might occur in big data & analytics projects, specifically a big data project in which several geographically-distributed teams from different organizations and backgrounds participate.

This paper is organized as follows. Section 1 presents the motivation for this research, the research objective, research questions and the contribution of this paper. In the section 2 the authors define the required concepts to understand the research. Section 3 presents the related work with respect to team coordination and big data process methodologies for project management. Section 4 presents the research method undertaken in this study. Section 5 presents the use case for the ASUM-DM based process methodology and the suggested prototype for a process methodology, which are the core of the authors' research. Finally, section 6 presents the conclusions and future directions in relation with the authors' research.

1.1 Motivation

According to Saltz et al., there is no accepted standard for a big data & analytics process methodology so far [20], therefore the work teams participating in a big data & analytics project often have to create an ad hoc methodology to handle the work dynamics within each team and among teams. The latter shows a low process maturity level [2] and might cause poor coordination and, consequently, the failure of the project. At the beginning of the reviewed big data & analytics project the designated work teams worked using an ad hoc methodology in an isolated way. Nevertheless, without a clear methodology to handle team work dynamics, integration and coordination several problems started to appear:

1. **Isolated research groups and corporate areas**. The research groups and corporate areas started to operate as isolated units without knowing what the other groups were doing.
2. **Communication issues**. There were situations where a team made an important decision and this decision was not communicated to all the teams or were only known by only some members of a team.
3. **Mistrust**. The teams did not know each other and so their research topics, objectives and interests. There was a generalized feeling of mistrust towards the other work teams. This hindered the cooperation among teams and the knowledge transfer.
4. **Unclear workflows**. There was no defined workflow. Thus, the teams did not have clear what inputs to expect from other teams and what outcomes the other teams were expecting from them.
5. **Different definitions and points of view about big data**. The work teams participating on the project had different points of view about what big data and big data & analytics were. This lead to delays in the implementation and the delivery of outcomes.

1.2 Research objective

Propose an implementation style for big data & analytics projects, whose main characteristic is that the work teams are cross-disciplinary and are geographically located far away from each other.

1.3 Research questions

- **RQ1**. How to integrate and coordinate several cross-disciplinary geographically-distributed teams belonging to different organizations to the success of a big data & analytics project in an experimental stage?
- **RQ2**. How can a workflow be defined for different research groups and corporate areas to achieve a big data project's goals?
- **RQ3**. Which criteria have to be taken in account to classify a data analytics project as a big data & analytics project?

1.4 Contribution

This proposal extended the External ASUM-DM process methodology for the implementation of big data & analytics projects. The use of the first phases of the methodology along with an analysis of the 5 V's of big data [6], the incorporation of a Business Process Model and Notation approach (BPMn) [14] and a strong communications management strategy was found to be the best solution for the coordination and communication of the different knowledge areas integrating a cross- disciplinary research-oriented big data project.

2 Concepts

Big Data. So far, there is no accepted definition for big data. However, the authors are going to use the definition proposed by Jin et al. Big data refers to a "bond that connects and integrates the physical world, the human society, and cyberspace in a subtle way" and can be classified into two categories, concretely, data from the physical world (e.g. sensors, scientific experiments and observations) and data from the human society, which (e.g. social networks, Internet, health, finance, economics and transportation) [10].

Big data's 5 V's Model Big data has certain characteristics also known as the 5 V's of big data:

1. *Volume* - high volume - refers to the size of the data. Big data sizes are given by multiple terabytes and petabytes.
2. *Variety* - high variety - refers to the different structural formats in a dataset. The data can be unstructured (e.g. text, audio, images, video), semi-structured (e.g. XML files) or structured (tabular data).
3. *Velocity* - high velocity - "refers to the rate at which data are generated and the speed at which it should be analyzed and acted upon".
4. *Veracity* represents the unreliability inherent in some data sources. Big data is often uncertain and imprecise.
5. *Value* refers to the real value which can be obtained from analyzing big data. The purpose of big data is to obtain a high value from its analysis of large volumes of data [6].

Analytics. The leading international association for professionals in analytics, The Institute for Operations Research and Management Sciences (INFORMS), defines analytics as "the scientific process of transforming data into insights for making better decisions" [15, 16].

Big Data Analytics. Big Data Analytics (BDA) refers to doing the analytics work directly in a big data environment. This activity requires more technical talent and programming skills than the traditional analytics [5].

CRISP-DM. The *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) is a data mining model created by IBM and Daimler AG in the late 1990s. This model encourages best practices and offers organizations the structure required to get better and faster results from data mining [21]. Despite its benefits, CRISP-DM lacks templates and guidelines and is weak on other activities that are also necessary for the success of big data projects, namely, infrastructure/operations, project management and deployment [9].

ASUM-DM. *The Analytics Solutions Unified Method for Data Mining/predictive analytics* (ASUM-DM) is an extended and refined version of CRISP-DM for implementing data mining and predictive analytics projects, which was created by IBM in 2015. ASUM-DM tried to compensate the weaknesses of CRISP-DM by adding new activities, incorporating templates and guidelines, and enhancing existing activities. There are currently two versions of ASUM-DM: an external version, which is offered in the web for free, and a proprietary version used by IBM internally [8]. This proposal is built upon the external version of ASUM-DM.

3 Related work

Kaskade [11] surveyed 300 organizations and found out that 55% of the big data projects are not completed, 39% of the times because of lack of cooperation. A poll conducted in 2014 by Piatetsky [17] showed that CRISP-DM remains the most popular methodology for data mining and data science projects. However, Saltz et al. [20] identified that the usage of CRISP-DM and SEMMA is decreasing and [1] pointed out that this is, in part, because the existing process methodologies do not consider knowledge management, communication management and project management.

Moyle [13] presents a CRISP-DM based Data Mining framework as well as guidelines for undertaking Collaborative Data Mining as a solution to the communication, coordination, cultural and social difficulties that resulted from the collaboration among geographically distributed laboratories and organizations by using collaborative internet-based tools. Also, Espinoza and Armour [5] identified that building big data & analytics abilities is not enough to successfully carry out a big data project, but coordination and governance also play a major role on this purpose. On the other hand, Grady [7] identified that mission expertise, domain data and processes, statistics, software systems and engineering, analytic systems and research and algorithms are the required skills that a big data team needs to successfully undertake a big data project. Bhardwaj [2] identified that collaborative data analysis and data science is often done following and ad hoc methodology and by doing trial-and-error. The latter shows a necessity for process maturity to guarantee the collaboration and coordination inside and between teams. Saltz [19] found out, that having a big data process methodology incorporating sprints, phases or some other set of well defined processes helps understanding the roles and allowing coordination in and between teams.

4 Research Method

To validate the proposal, the authors used an adaptation of the case study approach proposed by Runeson and Brereton [3, 18]. The tasks of this adaptation are Design, Data Collection, Data Analysis, Interpretation, and Reporting. This

tasks allowed the authors to research the adaptation of the ASUM-DM process methodology within a realistic context of a big data & analytics project.

Use Case. Bancolombia Group required to find a more precise measure of its clients' risk and profitability for strategic decision making within its corporate divisions. To achieve this goal the bank started a big data & analytics project named *Bancolombia Communities*. Bancolombia joined Eafit University and Icesi University, and together they sketched a solution consisting in a big data & analytics and network analysis approach that identified relevant communities of clients in function of their associated risk and profitability. *Bancolombia Communities* required specialized knowledge in statistics, economics, data visualization, applied mathematics, information technologies and project management. However, this knowledge was spread across several corporate divisions within the bank and research departments within the universities. Moreover, these corporate units and research departments were geographically-distributed in two distant cities (Medellin and Cali). This made the integration, communication and coordination among teams difficult and a key issue for the project's success.

The work team formation for the *Bancolombia Communities* project is explained as follows. Bancolombia's marketing department contributed with one work team consisting in two people: one person that was both domain expert and data scientist, and one data analyst. Eafit University counted on the participation of its school of engineering, school of sciences, and school of economics and finances. Each school contributed with a work team. The school of engineering contributed with one big data architect and IT infrastructure specialist, one knowledge management and database specialist, and three data engineers. The school of economics and finances contributed with one economics specialist, one statistics specialist, one economist and two finances analysts. The school of sciences contributed with two data scientists and one applied mathematician. Likewise, Icesi University counted on the participation of its design department, IT department, and accounts and finances department. The design department counted on the participation of one design and user experience specialist. The IT department counted on the participation of one data scientist, and one software engineer. The accounts and finances department counted on the participation of one economics specialist.

4.1 Use case design

The type of study used in the research was a primary qualitative study *in vivo* based on direct observation.

After data collection and analysis, the authors identified two stages for this case study. At the end of each stage the authors collected the generated information.

- **As-is stage**. The first stage lasted for 6 months. In this stage, all the work teams worked with their own ad hoc methodology and there was no knowledge of process methodologies for big data & analytics projects.
- **To-be stage**. The second stage lasted for one year and corresponded to the milestone, when the participants started to incorporate the proposed ASUM-DM process methodology.

The analysis at the end of each stage allowed the authors to establish a comparison point, where they could determine what influence the proposed process methodology had in the teamwork dynamics.

4.2 Use case data collection

The authors' data sources consisted in *in situ* observation of the daily activities, meeting reports, and both in-person and virtual meetings. Additionally, the authors reviewed all the generated work artifacts made by the work teams during one year, from March 2016 to October 2017, such as notes, technical reports, procedural reports etc.

4.3 Use case data analysis

The authors analyzed the collected data and identified the activities undertaken by the work teams during both the As-is and To-be stages.

Likewise, the authors mapped the developed activities with the ones proposed by ASUM-DM to recognize which activities were necessary and which not. The result of this analysis is an adaptation of the first phases of ASUM-DM - Analyze, Design, Configure, Build along with Project Management -.

5 Use case: ASUM-DM-based process methodology

Figure 1 shows the proposed workflow for the ASUM-DM adaptation. Each shape corresponds to one activity and the arrows correspond to loops among activities. The activities proposed by the authors are depicted in blue. These activities are, namely, (3.2) "Describe data against big data's 5 V's", (4) "Build prototype", (4.1) "Define prototype workflow", (4.5) "Build visualization" and (5) "Evaluate prototype". The rest of the activities are represented by white shapes and already existed in the original ASUM-DM process methodology. The blue arrow corresponds to a loop proposed by the authors. The black arrows, on the other hand, correspond to already existing loops in the original ASUM-DM process methodology [9].

The main differences between the original ASUM-DM and the proposed adaptation are the introduction of the activity (3.2) "Describe data against big data's 5 V's" to classify, at an early phase, whether the project is about big data. Another main difference is the introduction of the activity (4) "Build prototype". This activity was introduced because the proposal focuses on having a functional

big data application prototype and not on only having a valid and validated model. Since the scope of this article is the adaptation of a process methodology for big data & analytics projects, an expected final outcome would be a functional big data application prototype, which may be developed from scratch or be based upon existing commercial software. Additionally, the authors introduced the activity (5) "Evaluate Prototype" because, besides evaluating a model, the evaluation of the functional application prototype is necessary to determine how well it satisfies the business success criteria stated in the activity (2) "Understand Business". For this reason the authors also introduced a loop between the activity (5) "Evaluate model" and (2) "Understand business": if the prototype does not satisfy the business success criteria stated in the latter activity, the process has to be repeated from the activity (2) "Understand Business" up to the activity (5) "Evaluate prototype", until the prototype satisfies the business success criteria.

The rest of the loops follow the same pattern: they are repeated if the activity corresponding to the start of the arrow has not been successfully accomplished with respect to the business success criteria. There is also the case where the model built does not fit the prepared data and the people involved with the activity have to return to the (4) "Prepare Data" activity.

Furthermore, in the activity (4.1) "Define Prototype Workflow", the authors proposed the creation of a workflow by using a BPMn notation [14]. This activity is proposed to model the development process of the functional application prototype and, with that, guide the work teams on the expected inputs, outputs, and activities required for the construction of the application prototype.

5.1 Understand Data

Data sources. The complete clients and transactions dataset is stored in a traditional Relational Database Management System (RDBMS) and has more than 100,000,000,000 records containing all the clients of the bank and their transactions with other clients. The clients can be either personal or corporate clients. Bancolombia sampled this dataset with a non-random sampling criterion: selecting only the corporate clients (ca. 30,000). The purpose of this sampling was to start gaining useful insights at an early stage from the clients, from which the bank gets the most profit: its corporate clients. This is due to the fact that the corporate clients are the clients that carry out the transactions with the highest amount of money and move more money through the bank. Nonetheless, the processing and analytics of Bancolombia's complete clients and transactions dataset is left as a further phase and is not covered in this paper. Bancolombia sampled the input data and saved it into seven text files in the form of CSV. This was the input data set for the project. There was a total of seven CSV files analyzed with a total size of 3GB. At the end of this activity, the project management team elaborated a data understanding report.

Describe Data against Big Data's 5 V's. This proposal suggests an analysis of the collected data against the 5 V's identified by Gandomi et al. [6]: Volume,

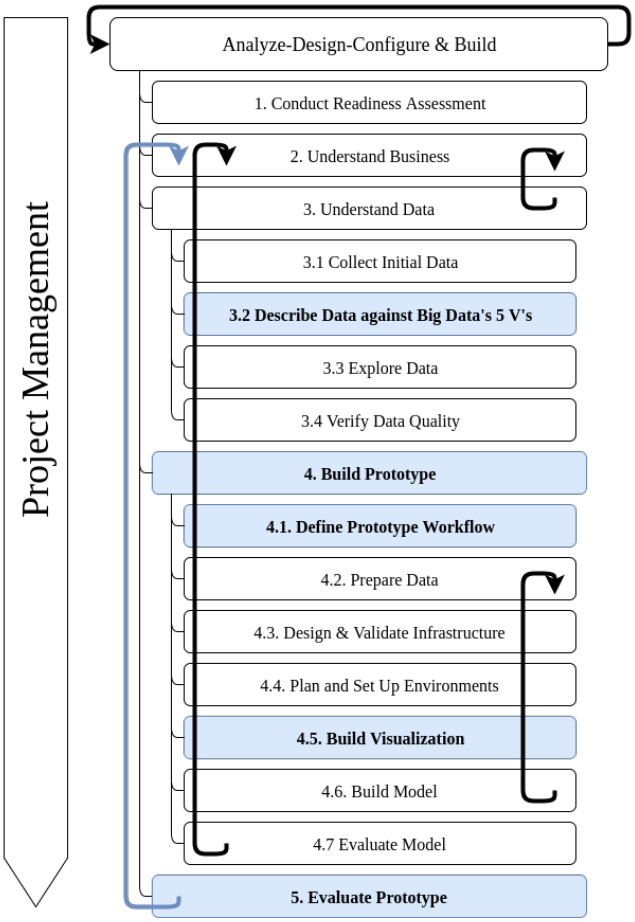

Fig. 1. Proposed adaptation of the ASUM-DM methodology

Variety, Velocity, Veracity and Value. The purpose of this activity is to identify the challenges inherent to the collected data, business requirements and business success criteria. With this information, the work team can easily identify the data challenges for the project in terms of big data and assess whether the project is about big data.

5.2 Build Prototype

This activity involves the construction of a functional application prototype that fulfills the business success criteria rather than a single model as proposed by ASUM-DM [9].

Define Prototype Workflow. The authors propose the modelling of an application prototype using a Business Process Model and Notation (BPMn) approach [14]. The project management team (Eafit University's school of engineering) found this approach to be easily understandable among all the different corporate departments and research groups. Thus, the activities to be carried out by each area for the prototype development were modelled as a Business Process. To solve the problems presented in the Motivation section, the authors carried out a series of meetings with each of the work teams to understand and model the activities that each team was undertaking and, hence, understand the responsibilities of each team, what each team received as input and what outputs a team should give to other teams.

Build Visualization. Eafit University and Icesi University created a visualization application based on graph plotting and dashboard design for the detection of risk and value communities.

5.3 Evaluate Prototype

In this activity, a functional application prototype is evaluated against the business objectives and business success criteria rather than a single model. With this activity, the stakeholders may identify if a potential big data & analytics application would fulfill its business necessities. The work team in charge for this evaluation was Bancolombia's marketing department.

5.4 Project Management

One of the key activities for the project management was the joint creation of the project plan. This activity allowed the project management team to record the common business understanding, to determine the business objectives and goals and to identify and assess the available resources, associated risks, constraints, and methodologies to use in the project. The work included from three to eight researchers in complementary disciplines for each corporate division and research group. To consolidate the confidence and relationships within each team, each work team held weekly in-person meetings. Likewise, to consolidate the confidence and relationships among teams from the different universities and cities, the teams held virtual meetings. The purpose of those meetings was to synchronize the work developed by each research team. Moreover, the work teams undertook quarterly in-person workshops to improve the interactions among teams and to ease the knowledge transfer among the three organizations. Undoubtedly, the technical committee support (Eafit University's school of engineering) among the researchers from different universities was fundamental, because it allowed to share challenges found and learned lessons by the work teams about the projects carried out under the CAOBA Alliance. From a technical point of view, the project management team established some minimal configuration

management practices to support the engineering activities and project management within the project. The project management team built a document management system incrementally to have a register and a traceability for the decisions made in the meetings and workshops undertaken throughout the entire project. Also, all the technology products generated by the different organizations, such as software applications and process methodologies was saved in this document management system.

6 Conclusion and Future Work

With the ASUM-DM-adapted big data process methodology for multiorganization, multidisciplinary, geographically distributed teams, it was possible to define a solution that supported the necessities of a big data & analytics project in the project management dimension. Those necessities were not explicitly defined by the ASUM-DM activities themselves, because the ASUM-DM methodology addresses neither the problem of a big data project, nor the context of cross-disciplinary, geographically-distributed and multi-organization research groups and corporate divisions.

Likewise, this adaptation generated a teamwork dynamic that improved the communication and team collaboration through the organization around a workflow and the help of communication policies. The authors think that this adaptation will set a start point for researchers and practitioners when choosing or tayloring a process methodology for the experimental phase of a big data & analytics project in the context previously described.

A further work for this study might be the evaluation of the proposal by carrying out surveys and focus groups to researchers and practitioners participating in big data projects. Also a study about similar data analytics projects might be carried out with the aim of knowing the challenges and learned lessons when adopting this adapted process methodology as a base methodology for the implementation of projects in a similar context.

Acknowledgment

The authors would like to thank the *Colombian center of excellence and Appropriation On Big data & data Analytics* (CAOBA Alliance) for providing the funds for this study. Also, the authors would like to thank Bancolombia group, Eafit University and Icesi University, specially professors Diego Restrepo, and Juan Manuel Salamanca for their collaboration. Finally, the authors would like to thank the Colombian Administrative Department of Science Technology & Innovation (COLCIENCIAS), and the Colombian Ministry of ICT (MINTIC), both members of the CAOBA Alliance.

References

1. Ahangama, S., Choon, C., Poo, D.: Improving Health Analytic Process through Project , Communication and Knowledge Management. Icis-Rp pp. 1–10 (2015)

2. Bhardwaj, A., Bhattacherjee, S., Chavan, A., Deshpande, A., Elmore, A.J., Madden, S., Parameswaran, A.G.: DataHub: Collaborative Data Science & Dataset Version Management at Scale (2014), http://arxiv.org/abs/1409.0798
3. Brereton, P.: Using a Protocol Template for Case Study Planning (2006)
4. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: Crisp-Dm 1.0. CRISP-DM Consortium p. 76 (2000)
5. Espinosa, J.A., Armour, F.: The big data analytics gold rush: A research framework for coordination and governance. Proceedings of the Annual Hawaii International Conference on System Sciences 2016-March, 1112–1121 (2016)
6. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35(2), 137–144 (2015), http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007
7. Grady, N.W.: Knowledge Discovery in Data Science. KDD meets Big Data pp. 1603–1608 (2016)
8. Haffar, J.: Have you seen ASUM-DM? (2015), https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/
9. IBM: IBM Analytics Solutions Unified Method (ASUM) (2015)
10. Jin, X., Wah, B.W., Cheng, X., Wang, Y.: Significance and Challenges of Big Data Research. Big Data Research 2(2), 59–64 (2015), http://dx.doi.org/10.1016/j.bdr.2015.01.006
11. Kaskade, J.: Cios big data: What it teams want their cios to know (2013), http://blog.infochimps.com/2013/01/24/cios-big-data/
12. Matignon, R.: Data Mining Using SAS® Enterprise Miner (2007)
13. Moyle, S.: Collaborative Data Mining. No. 54, 2 edn. (2010)
14. Object Management Group (OMG): Business Process Model and Notation (BPMN) Version 2.0. Business 50(January), 170 (2011), http://books.google.com/books?id=GjmLqXNYFS4C&pgis=1
15. for Operations Research, T.I., the Management Sciences (INFORMS): About informs (2017), https://www.informs.org/About-INFORMS (Visited in November 2017)
16. for Operations Research, T.I., the Management Sciences (INFORMS): Operations research & analytics (2017), https://www.informs.org/Explore/Operations-Research-Analytics (Visited in November 2017)
17. Piatetsky, G.: Crisp-dm, still the top methodology for analytics, data mining, or data science projects (2014), http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html
18. Runeson, P., Martin, H., Rainer, A., Regnell, B.: Case study research in software engineering :Guidelines and Examples (2012)
19. Saltz, J.S.: The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015 pp. 2066–2071 (2015)
20. Saltz, J.S., Shamshurin, I.: Big Data Team Process Methodologies : A Literature Review and the Identification of Key Factors for a Project ' s Success pp. 2872–2879 (2016)
21. Shearer, C., Watson, H.J., Grecich, D.G., Moss, L., Adelman, S., Hammer, K., Herdlein, S.a.: The CRISP-DM model: The New Blueprint for Data Mining. Journal of Data Warehousing14 5(4), 13–22 (2000), www.spss.com%5Cnwww.dw-institute.com