Contribution in Big Data Projects Management

Abderrachid Errezgouny, [1, *], and *Abdeljabbar* Cherkaoui[1]

[1]Laboratory of Innovative Technologies (LIT), National School of Applied Sciences,
Abdelmalek Essaadi University, Tangier, Morocco

Abstract. Nowadays, the necessity of the data becomes more attractive by companies in different areas (IT, space, automotive) who need to create and capture the value from the huge amounts of data generated from various sources. Many fields need to use this amount in the right way in real time with high level processing, this evolution is called Big Data (BD). In this case, to manage a BD project the specific tools like Machine Learning, Data Mining, and more are very important to achieve the customer satisfaction with the expected quality of services. The majority of BD projects fall due to the lack of managing skills and team training, also the sophisticated materials and technologies are required. This paper presents our contribution in the project management of BD based on other discussed methods like Project Management Body of Knowledge (PMBoK) and Agile approaches, and we use them to construct a rigid model for managing any project dedicated to work with BD.

1 Introduction

In general, the project management (PM) turn around three terms (Time, Scope, Cost) for assuring good quality for products or services development and customers satisfaction. Companies in different areas needs to digitalize their process by working with huge amounts of data at scale to enhance the quality and rapid response. In this case, the right project management approach specified to BD is required.

Companies are realizing a great potential that BD technologies can bring to improve their performance, businesses, and to increase their competitive advantage [1]. In any areas the notion of BD is usually common, and the move towards the digitalization is more crucial for the majority of companies who seek to save her brand-name. The majority of organizations need to scale their business and development, the necessity of a flexible architecture includes security and agility is the most critical requirement for a BD project. Otherwise, 55% of BD projects don't get completed, and many others fall short of their objectives due to enormous reasons like, lack of business context around the data and lack of expertise [2]. There are many others reasons whose might guaranteed fail of BD projects, but with a robust model management and flexible architecture, we can transform the challenges to the opportunities. Other, through an exploratory study and detailed BD analyses based on 325 responses, the majority of 70% considers BD as an opportunity by using this massive information for business advantage for example a user organization can discover new facts about their customers, markets, partners, costs, and operations. [3]. To use this notion in the right way, you need to understand your data and apply some helpful technical like ML and DM, and to protect your data from cyberattack and hackers. In [4], present some BD applications which are customer need identification, creation product and service design, decision-making, data-driven knowledge, risk management, quality management, and opportunity recognition. In this context we focus on application of BD project management and assuring reliability and customers satisfaction, based on the traditional PM and Agile approaches. The presented work attempts to answer the following questions;

- What approach is suitable for the management of BD projects?[5]
- What are the keys factors of success the management of BD projects?

The rest of the paper is organized as follows, section 2 reviews the related work. Section 3 deals with the comparative view between PMBoK guide and Agile in BD management, whereas section 4 presents the proposed model. Finally, we conclude our work in section 5.

2 State of art

2.1. Big Data background

Data increase more and more in various areas and applications, due to the increasing of the number of companies who search for digitalization, according to [6]. The BD has a big impact in different areas like the e-commerce and marketing intelligence (recommendation management systems, social media monitoring and analysis), government and politics, science and

[*] Corresponding author: abderrachid.errezgouny@etu.uae.ac.ma

technology (innovation, code errors verification, hypothesis testing, and knowledge discovery), smart health and wellbeing and recently in e-learning. And also, in the Intelligent Transportation Systems (ITS) BD applications including object detection, data storage and archiving, traffic flow prediction, signal recognition, and finding the optimal route and safety of the vehicle and road [7]. The widespread use of data is increasing, which is why the traditional methods of collecting and analyzing data aren't compatible with BD projects. This term defined by [8], as the amount of data exceeds the effective storage, management and processing capabilities of the technology. Seagate company predict the growth of 30% of the world's data will need to be processed in real time (from 45 Zettabytes in 2019 to 175 Zettabytes by 2025) [9]. The BD characterize by three dimensions (3V's) originally proposed by Doug Laney in 2001[10], are i) "Volume" represents the primary attribute include terabytes, records, transactions, tables and files, ii) "Variety" where data are coming from various sources than even before, and iii) "Velocity" which is the speed and the frequency of data generation or data delivery, in our case we will be focusing on streaming data. There are more characteristics presented by [8] are data value and complexity, those measures respectively the usefulness of data in decision-making, and the degree of interconnectivity. And in term of variety, we need to understand our data' type, structured, semi-structured, and unstructured information.

With this evolution, the application and the process are going more complicated. The main BD challenges is how to capture, collect, manage, analyze and to be able to understand the sense of data and to take actions, like decision making, live dashboard and reporting. Data volume and variety sources can be representing a challenge for companies. Another challenge is presented by [8], is the dynamic design for the end-users, i.e., the system' designer requires to consider the technologies and the needs of users. To address the enormous challenge in BD a specific model will be needed to manage and control the project more efficiently.

2.2 PMBoK and Agile methods to manage Big Data projects

2.2.1 PMBoK Guide

Is defined by PMI as a term to describe the knowledge within the profession of PM, including proven traditional practices used, and also different innovative skills that are emerging in the profession [11]. This approach categorized by five process groups (initiating, planning, execution, monitoring/controlling and closing), and ten knowledge areas adapted with BD project [11], [12];

- Project Integration Management. The integration of the BD technology with existing tools and techniques in company is also an issue related to the project that should be solved within Integration Management processes [13]. This integration should take in advance any tools related to the BD like technology used for gathering information, data processing, management of databases, servers, etc.
- Project Scope Management. Including processes required to ensure if the project contains all tasks, documents, to complete the project successfully. Data collection and data preparation should include in the scope of the BD project [13].
- Project Schedule Management. This process used to track the timely completion of the project tasks. In BD projects we need to take on consideration many points in the scheduling e.g., tasks complexity, changes and modifications.
- Project Cost Management (PCM). This phase includes all processes to control and manage the cost of the whole project like scheduling, estimating, financing, budgeting, etc. In addition, BD projects usually involve hidden costs and complexity, which makes the planning and control process more difficult [13].
- Project Quality Management. Assuring the data quality by respecting the quality policy and procedures regarding planning, managing, and controlling the BD project to achieve the product quality requirements and the best performance, in order to meet stakeholders' expectations.
- Project Resource Management. It is the same case comparing it to any kind of project, the team work should have the required skills with different background and expertise like a data scientist, data architect, software engineers, data technicians, etc. And also, by assuring the effective and the efficient technologies to achieve higher results.
- Project Communication Management. According to PMI, the effective communication is the most crucial factor in a BD project.
- Project Risk Management. This phase includes processes of controlling and conducting the risk. In terms of planning, identification, risk analyzing, response planning, implementation and monitoring the risk. Some additional risk associated with BD such as cyber-attacks, BD architecture, lack of knowhow in the data filed [13].
- Project Procurement Management. It includes all necessary processes of purchasing new services or products (materials), e.g., the decision to move to the cloud, the company will take on consideration the vendors' problems, license management, and the marketing strategy.
- Project Stakeholder Management. It is the most important part in our development, that to involve the stakeholder (people, groups, or organizations) in BD project to augment the efficient rate, by analyzing the expectations, engaging, and their impact on the project.

For a specific project type, the needs of additional knowledge areas may be required (construction, aerospace, etc.). The traditional approach is based on a sequence of project management process steps as explained in PMBoK by PMI [11], (See Figure 1).

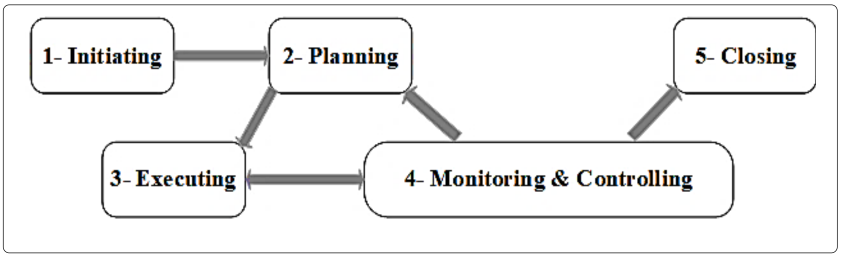

Figure 1. The five process groups of the PMBoK Project Management process.

The traditional project management assumes that events are predictable and techniques and tools are well understood by companies [14]. Some strengths of PMBoK are well-structured process, easily teachable and repeatable, and it gives the im-portance to requirements. The application of this approach is widely used in the industry for example to develop new product and technology, technical support tools, and process improvement and sustainability strategies. This approach in BD management needs to be dynamic and flexible with any changes in the later project stages.

2.2.2 Agile Project Management (APM)

At the beginning, Agile manifesto was released in 2001 by seventeen software process methodologists, who create four values for Agile software development for better way of developing software; Individuals and interactions over processes and tools, working software over comprehensive documentation, customer collaboration over contract negotiation, and responding to change over following a plan [15]. In selection phase, the project teams need awareness of the options and tools availability to select the most likely approach to be successful for the project. APM which is mainly used in software development and adapted for another kind of project. In [16] demonstrate in his exploratory study by surveyed 19 Brazilian group of companies, that APM could be adapted to non-software companies, or more traditional industry sectors, and for a part of a project. APM is a highly iterative and incremental process, where developers and project stakeholders actively work together to understand the domain, identify what needs to be built, and prioritize functionality [14].

Agile is not only a method but also a mindset, it's preferably to use it for projects with less rigid constraints, smaller risks, and in the collocated environment. There are many software development methods include in Agile like, Scrum wish used for a small team [17], Extreme Programming (XP) characterized by its simplicity form and fives practices areas (organizational, technical, planning, and integration), and Feature-driven development (FDD) is an iterative and incremental software development process providing a flexibility of requirements changing at the end of the project. Each of them defining by its own processes or techniques for realizing the core principles for agile methods [18].

3 Comparative study

The comparison between PMBoK and Agile methods, resume in the answers of these two following questions; how a project is to be carried out? What should be done? In BD project management is the same case, that we seek always the answers. But the complexity is the adaptation of the discussed methods or other methods to the BD project. We can say the PMBoK is the general case of Agile. [19] Present in his work a number of comparisons between the nine knowledge areas and Agile methods process (XP, Scrum, and FDD), the conclusion of this study is that PMBoK is an exhaustive list of good practices that can be customized to specific needs. Otherwise, Agile methods do not define all facets needed in PM and doesn't fully address some knowledge areas like risk, cost and procurement management. [18] and [20], present some challenges related to Agile e.g., lack of experience, insufficient training, project complexity, etc. Utilizing Agile methodologies have many advantages for organization e.g., quick changes, flexibility, regular business development environment and rise of customers' expectations by decreasing the time delivery [21].

We can use the hybridization of approaches like [20], used a Waterfall process and APM for profiting more agility and flexibility. This term is widely used in many manufacturing structures by combining two or more methods to gain advantages. The reason of companies being so actively experimenting and adopting multiple approaches is to deliver successful projects to maximize profits and Return on Investment (ROI) [20].

There is different approach to manage a project, but Agile is widely used in data environment, especially with software development. In the rest of this paper, we are going to propose our scalable model based on both discussed approaches.

4 The proposed model and discussions

The term of the management in BD project carried out many definitions, what we manage exactly? Team, data or both. The integration of the management skills in the BD project necessary to follow some steps, first of all we need to know the scope of the project and know our data type. We used the flexibility and the simplicity of agile approaches and the robust standard principals of PMBoK to develop our BD model, which takes on consideration two principal aspects; the team work' management and the technical view. Summarized in the following six steps model (See Figure 2):

Initiating. This is a phase where a BD project is set up, we start to understand our project with the collaborative team, which involve defining the scope of the project by creating technical documentations. The construction of the teamwork is more critical, by choosing the team who has collaborative manner adapted in different situations like the tele-work, and with the required skills.

Planning and Modelling. This is the phase when we schedule our work by dividing the whole BD project into small tasks assigned to each member or small team, and defining key milestones to understand what we need in term of the project' requirements to get completed. Then, we start to collect the data from various BD sources (Internet of Things, data centers,) by using historical information about the event, the web logs, and other technologies. This step depends on the kind of the project, that respecting the specifications of the customer.

Execution. Here, we are going to release the previously planned, after collecting the data, the cleaning and treatment are considered as a key stage. Data cleaning eliminates the incorrect values (noise errors, redundancies, incomplete data, outliers, etc.) and checks for data inconsistency [22]. And [23], represent data cleaning as a phase in data preprocessing.

Implementation and Test. In this phase, the team can release a prototype of the project and test them in the practical environment according to the standard of work provided by the customer before the validation process.

Controlling. A BD project must be monitored and controlled by factors like if the project is respecting the budget, the level of achievement the task scheduled and by checking the quality targeted and resources. In this phase, the modification changes from the customer are acceptable, and the team needs to be flexible and adapt to any situation by preparing preliminarily a plan B. To ensure service quality, the team generates a performance report describing the result and the technical view.

Data usage and Visualization. This last phase involves the final inspections and reliability tests to ensure that the expected results are achieved. There are many traditional ways to visualize the data, e.g., bar charts, pie charts, etc. according to the Forrester Wave report, present the Advanced Data Visualization (ADV), like a modern technology using more interactive and dynamic business graphics, such as live synchronous, real-time dashboards and charts that update automatically with the data state [24]. The most helpful software Microsoft Power BI, and Sisense that visualize any amount and any type of data at real time.

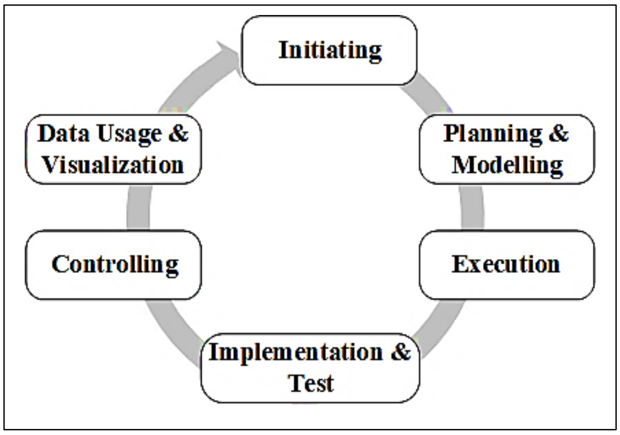

Figure 2. The proposal model for Big Data Project Management

Finally, the customer will be able to use his net data fluidly and make his own strategic decision. The output of this model can be reused like an input of others projects, recycling of the data, etc. By following this model, we attend to achieve high productivity and performance level and many success factors in term of the ability to store and access the appropriate data, governance with well defining roles and responsibilities, rigid process thanks to formal methodologies, objectives achieving, developing the team' skills in data-driven decision-making, and tools insights [25].

5 Conclusions and futures work

The complexity of the managing Big Data projects is in the scalability and the integration in different sectors of the data generated and processing. The comparison between two formal approaches shows that Agile and PMBoK share the same mind-set which highly used in traditional projects, with the lack of flexibility required, stakeholders' integration in the whole process development, and the lack of the compatibility with BD.

Hereinafter, the move to the next generation of networking and the digitalization requires more attention to the data generated and how to manage it in any firms. Our proposed model is not the perfect solution, but maybe the best thing to do before all steps toward the implementing of the new digital system. That gives the possibility of scaling the project as we need and provides more flexibility than Agile and other traditional approaches. To choose the best approach to manage BD is a decision depends on the company and other factors. This proposition is still needing development and adaptation to validate in the environment and test the reliability at high scale of volume, variety and complexity of data. This BD area requires further investigation, research, by integrating others technologies like native cloud, using machine learning and data mining to enhance the data processing.

References

1. C. Ponsard, M. Touzani, and A. Majchrowski, "*How to conduct big data projects: Methods overview and industrial feedback*," Ing. des Syst. d'Information, vol. **23**, no. 1, pp. 9–33, (2018), doi: 10.3166/ISI.23.1.9-33.
2. J. Kelly, J., & Kaskade, "*CIOs S & BIG DATA What Your It Team Wants You To Know*," (2013), doi: http://blog.infochimps.com/2013/01/24/cios-big-data/.
3. P. Russom, "*Big Data Analytics*," TDWI Res., vol. **38**, pp. 38–48, (2011), doi: 10.1017/9781108566506.005.
4. A. Urbinati, M. Bogers, V. Chiesa, and F. Frattini, "*Creating and capturing value from Big Data: A multiple-case study analysis of provider companies*," Technovation, vol. **84–85**, no. May 2018, pp. 21–36, (2019), doi: 10.1016/j.technovation.2018.07.004.
5. P. Franková, M. Drahošová, and P. Balco, "Agile Project Management Approach and its Use in Big Data Management," Procedia Comput. Sci., vol. **83**, no. Ant, pp. 576–583, (2016), doi : 10.1016/j.procs.2016.04.272.
6. H. Chen, R. H. L. Chiang, and V. C. Storey, "*Business Intelligence and Analytics: From Big Data to Big Impact*," MIS Q., vol. **36**, no. 4, pp. 1165–1188, (2012).
7. S. Kaffash, A. T. Nguyen, and J. Zhu, "*Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis*," Int. J. Prod. Econ., vol. **231**, no. April 2020, p. 107868, (2021), doi: 10.1016/j.ijpe.2020.107868.
8. S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "*Big data: Issues and challenges moving forward*," Proc. Annu. Hawaii Int. Conf. Syst. Sci., pp. 995–1004, (2013), doi : 10.1109/HICSS.2013.645.
9. "SEAGATE." https://www.seagate.com/ (accessed Jun. 20, 2021).
10. D. Laney, "*3D Data Management: Controlling Data Volume, Velocity, and Variety.,"* META Gr., p. 4.
11. *PMI, A Guide to the Project Management Body of Knowledge*, Sixth Edit. (2017).
12. H. Middleton et al., *Agile practice guide. Project Management Institute*, (2017).
13. A. A. Tokuç and Z. E. Uran, "*Management of Big Data Projects: PMI Approach for Success,*" PMI, pp. 279–293, (2019), doi: 10.4018/978-1-5225-7865-9.ch015.
14. K. B. Haas, "*The Blending of Traditional and Agile Project Management*," PM World Today, vol. **IX**, no. V**,** pp. 1–6, (2007).
15. "Manifesto for Agile Software Development," 2001. https://agilemanifesto.org/ (accessed Sep. 26, 2021).
16. E. C. Conforto, F. Salum, D. C. Amaral, S. L. da Silva, and L. F. M. de Almeida, "*Can Agile Project Management Be Adopted by Industries Other than Software Development?*" Proj. Manag. J., vol. **45**, No. **3**, no. July, (2014), doi : 10.1002/pmj.
17. L. Rising, N. S. Janoff, and a G. C. Systems, "*The Scrum Software Development Process for Small Teams*," Software, IEEE, vol. **17**, Issue, no. August, pp. 26–32, (2000), [**Online**]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=854065&tag=1.
18. G. J. Miller, "*Agile problems, challenges, & failures*," PMI® Glob. Congr., (2013), [**Online**]. Available: https://www.pmi.org/learning/library/agile-problems-challenges-failures-5869.
19. P. Fitsilis, "*Comparing PMBoK and agile project management software development processes*," Adv. Comput. Inf. Sci. Eng., pp. 378–383, (2008), doi : 10.1007/978-1-4020-8741-7_68.
20. A. Hassan, S. Younas, and A. Bhaumik, "*Exploring an agile plus approach for project scope, time, and cost management*," Int. J. Inf. Technol. Proj. Manag., vol. **11**, no. 2, pp. 72–89, 2020, doi : 10.4018/IJITPM.2020040105.
21. S. Abdalhamid and A. Mishra, "*Adopting of agile methods in software development organizations: Systematic mapping*," TEM J., vol. **6**, no. 4, pp. 817–825, (2017), doi: 10.18421/TEM64-22.
22. H. Mousannif, H. Sabah, Y. Douiji, and Y. O. Sayad, "*From big data to big projects: A step-by-step roadmap*," Proc. - Int. Conf. Futur. Internet Things Cloud, FiCloud, pp. 373–378, (2014), doi : 10.1109/FiCloud.2014.66.
23. S. B. Kotsiantis and D. Kanellopoulos, "*Data preprocessing for supervised leaning*," Int. J. …, vol. **1**, no. 2, pp. 1–7, (2006), doi: 10.1080/02331931003692557.
24. B. Evelson and K. TaKeaWays, "The Forrester WaveTM: *Advanced Data Visualization (ADV) Platforms*, Q3 (2012)," Forrester Res. Inc, [**Online**]. Available: www.forrester.com.
25. J. S. Saltz and I. Shamshurin, "*Big data team process methodologies: A literature review and the identification of key factors for a project's success*," Proc. - 2016 IEEE Int. Conf. Big Data, Big Data, pp. 2872–2879, (2016), doi: 10.1109/BigData.2016.7840936.