Process Guidance for the Successful Deployment of a Big Data Project: Lessons Learned from Industrial Cases

Christophe Ponsard[1], Annick Majchrowski[1], Stephane Mouton[1] and Mounir Touzani[2]
[1]*CETIC Research Centre, Charleroi, Belgium*
[2]*Académie de Toulouse, Toulouse, France*

Keywords: Big Data, Process Model, Agile, Adoption, Case Studies.

Abstract: Nowadays, in order to successfully run their business, companies are facing the challenge of processing ever increasing amounts of data coming from digital repositories, enterprise applications, sensors networks and mobile devices. Although a wide range of technical solutions are available to deal with those Big Data, many companies fail to deploy them because of management challenges and a lack of process maturity. This paper focuses on those aspects and reports about lessons learned when deploying a series of Big Data pilots in different domains. We provide feedback and some practical guidelines on how to organise and manage a project based on available methodologies, covering topics like requirements gathering, data understanding, iterative project execution, maturity stages, etc.

1 INTRODUCTION

Our world is currently experiencing an information explosion. Many figures are available to quantify the exponential rate of the Big Data phenomenon. For example, it is reported that 90% of the worlds data has been produced in just the last two years, and that the amount of data created by businesses doubles every 1,2 years (Rot, 2015).

Organizations typically view Big Data technologies as holding a lot of potential to improve their performance and create competitive advantage. The ease to collect and store data, combined with the availability of analysing technologies (such as NoSQL Databases, MapReduce, Hadoop) has encouraged many of them to launch Big Data projects. However most organisations are actually still failing to get business value out of their data. A 2013 report surveying 300 companies about Big Data revealed that 55% of Big Data projects dont get completed, and many others fall short of their objectives (Kelly and Kaskade, 2013). An on-line survey conducted in July 2016 by Gartner reported that many companies remain stuck at the pilot stage and that only 15% actually deployed their big data project to production" (Gartner, 2016).

Looking at the cause of such failures, it appears that the main factor is actually not the technical dimension but rather the process and people dimensions which are thus equally important (Gao et al., 2015). However, looking at the literature, the technical dimension is often emphasised - especially the use of algorithms that will produce a sharp analysis - while much less is devoted to methods and tools that can help teams to achieve big data projects more effectively and efficiently (Saltz and Shamshurin, 2016). There is however some recent work in that area, identifying key factors for a projects success (Saltz, 2015), stressing management issues (Corea, 2016), insisting on the need for team process methodologies and making a critical analysis of analytical methods (Saltz and Shamshurin, 2016).

Our paper is aligned with those works and aims at helping companies engaging in a Big Data adoption process to be driven by questions such as:

- How can we be sure Big Data will help us?
- Which people with what skills should be involved?
- What steps should be done first?
- Is my project on the right track?

Our contribution is of practical nature and composed of guidelines and lessons learned from a set of pilot projects covering various domains (life sciences, health, space, IT infrastructures). Those pilots are spread overs two years and are conducted in the scope of a common global project carried out in Belgium. They are following a similar process which is incrementally enhanced.

Ponsard, C., Majchrowski, A., Mouton, S. and Touzani, M.
Process Guidance for the Successful Deployment of a Big Data Project: Lessons Learned from Industrial Cases.
DOI: 10.5220/0006357403500355
In *Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security (IoTBDS 2017)*, pages 350-355
ISBN: 978-989-758-245-5

This paper is structured as follows. In Section 2, we review the main available methodologies for dealing with Big Data deployment. In Section 3, we present the process followed to come up with a method and validate it on our pilots. It stresses key requirements for successful deployment. Section 4 presents more detailed feedback and highlight specific guidelines. Finally section 5 draws some conclusions and possible extensions of our work.

2 SURVEY OF EXISTING METHODS AND PROCESSES

This section reviews existing methods and processes. It highlights some known strengths and limitations. First, methods inherited from the related data-mining field are presented before considering approaches more specific to Big Data with a special attention to Agile methods.

2.1 Methods Related to Data-mining

Data-mining developed in the 1990's with the aims to extract data patterns in structured information (databases) to discover business factors on a relatively small scale. In contrast, Big Data is also considering unstructured data and operates on a larger scale. However a common point from a process point of view is that both require the close cooperation of data scientists and management in order to be successful. Many methodologies and process models have been developed for data mining and knowledge discovery (Mariscal et al., 2010).

The seminal approach is KDD (Knowledge Discovery in Database). It was refined into many other approaches (like SEMMA, Two Crows, etc) before being standardised under CRISP-DM (Cross Industry Standard Process for Data Mining) (Shearer, 2000). This method is depicted in Figure 1. It is composed of six main phases each decomposed in sub-steps. The process is not linear but rather organised as a global cycle with usually a lot of back and forth within and between phases. CRISP-DM has been widely used for the past 20 years, not only for data-mining but also for predictive analytics and big data projects.

CRISP-DM and the like however suffer from the following issues:

- they fail to provide a good management view on communication, knowledge and project aspects.
- they lack some form of maturity model enabling to highlight more important steps and milestones that can be progressively raised.

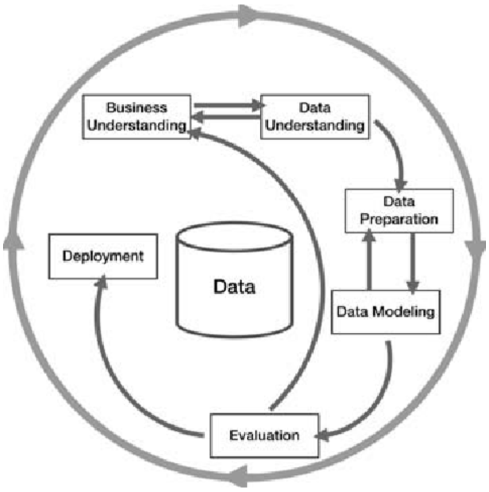

Figure 1: CRISP-DM Method.

- despite the standardisation, they are hardly known to the wider business community, hence difficult to adopt for managing the data value aspect.

2.2 Going the Agile Way

Agile methods, initially developed for software development, can be applied to data analysis in order to provide a better process guidance and value orientation. An agile evolution of KDD and CRISP-DM is AgileKDD (do Nascimento and de Oliveira, 2012). It is based on the OpenUP lifecycle which supports the statement in the Agile Manifesto (Balduino, 2007). Projects are divided in planned "sprints" with fixed deadlines, usually a few weeks. In each sprint, the teams need to deliver incremental value to stakeholders in a predictable and demonstrable manner.

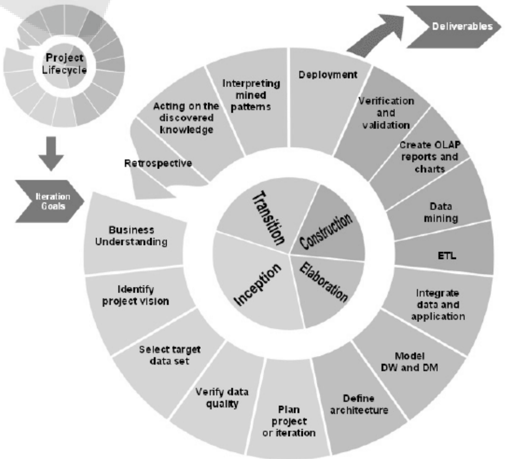

Figure 2: Agile KDD Method.

Although it looks quite adequate, deploying an Agile approach for Big Data may still face resistance, just as it is the case for software development, typically in more rigid kind of organisation. A survey was conducted to validate this acceptance (Frankov et al., 2016). It revealed that quite similarly as for software, companies tend to accept Agile methods for projects with smaller scope, less complexity, with few security issues and inside organisation with more freedom. Otherwise, the plan-managed approach is preferred.

2.3 Methods Specific to Big Data

Architecture-centric Agile Big data Analytics (AABA) addresses technical and organizational challenges of Big Data(Chen et al., 2016). Figure 3 shows it supports an agile delivery. It also integrates the Big Data system Design (BDD) method and Architecture-centric Agile Analytics with architecture-supported DevOps (AAA) model for effective value discovery and continuous delivery of value.

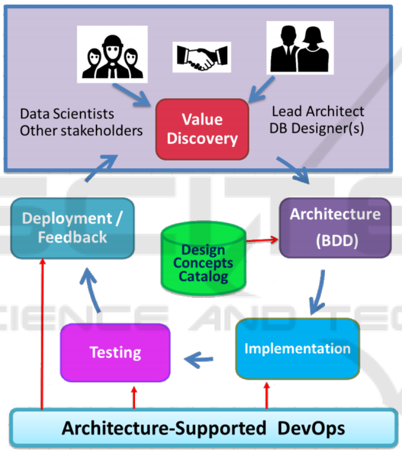

Figure 3: AABA Method.

The method was validated on 11 case studies across various domains (marketing, telecom, healthcare) with the following recommendations:

1. Data Analysts/Scientists should be involved early.
2. Continuous architecture support is required.
3. Agile bursts of effort help to cope with rapid technology changes and new requirements.
4. A reference architecture enables more flexibility.
5. Feedback loops need to be open, e.g. about non-functional requirements such as performance, availability and security, but also for business feedback about emerging requirements.

In parallel, Stampede is a method proposed by IBM to their customers where expert resources are provided at cost to help companies to get started with Big Data in the scope of a well-defined pilot project (IBM, 2013). It main goal is to educate companies and help them get started more quickly, in order to drive value from Big Data. A key tool of the method is a half day workshop to share definition, identify scope/big data/infrastructure, establish a plan and most importantly establish the business value. The pilot execution is typically spread over 12 weeks and carried out in an Agile way with a major milestone at about 9 weeks.

Some attempts have also been made to develop a Capability Maturity Model (CMM) for scientific data management practices, with the goal of supporting assessment and improvement of these practices (Crowston, 2010)(Nott, 2014). Such a model describes key process areas and practices necessary for effective management. A CMM further characterizes organizations by a maturity level, i.e. the capability to reliably perform the processes, typically on a 5 level scale (from the lowest just ”defined” or ”ad hoc” levels to the highest ”optimised” or ”Breakaway” level).

2.4 Complementary Approaches

Sensemaking is also an iterative approach but relating to the cognitive process performed by humans in order to build up a representation of an information space for achieving his/her goal. It focuses on challenges for modelling and analysis by bringing cognitive models into requirements engineering, in order to analyse the features of data and the details of user activities (Lau et al., 2014).

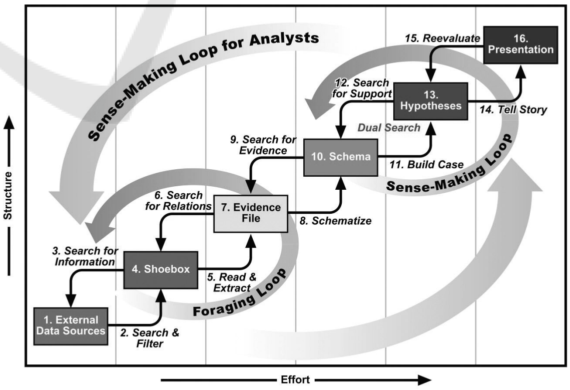

Figure 4: SenseMaking Method.

In complement to processes, many key success factors, best practices and risk check lists have been published, mostly in blogs for CIOs, e.g. (Bedos, 2015). A systematic classification of Critical Success Factors has been proposed by (Gao et al., 2015) using three key dimensions: people, process and technology. It has been further extended by (Saltz and

Shamshurin, 2016) with tool and governance dimensions. A few key factors are the following:

- Data: quality, security, level of structure in data
- Governance: management support, well-defined organisation, data-driven culture
- Objectives: business value identified (KPI), business case-driven, realistic project size
- Process: agility, change management, maturity, coping with data growth
- Team: data science skills, multidisciplinarity
- Tools: IT infrastructure, storage, data vizualisation capabilities, performance monitoring

3 METHOD DEVELOPMENT AND VALIDATION PROCESS

The global aim of our project is to come up with a systematic method to help companies facing big data challenges to validate the potential benefits of a big data solution. The global process is depicted in Figure 5, and is driven by eight successive pilots which are used to tune the method and make more technical bricks available through the proposed common infrastructure. The final expected result is to provide a commercial service to companies having such needs.

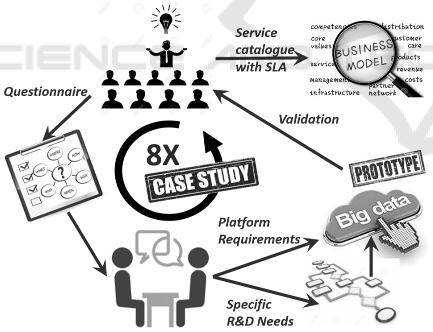

Figure 5: Iterative development of the platform and method.

The selected method is strongly inspired by what we learned from the available methods and processes described in Section 2:

- the starting point was Stampede because of some initial training and the underlying IBM platform. Key aspects kept from the methods are the initial workshop with all stakeholders, the realistic focus and a constant business value driver.
- however to cope with the lack of reference material, we defined a process model based on CRISP-DM which is extensively documented.
- the pilots are executed in an Agile way, given the expert availabilities (university researchers), the pilots are planned over longer periods than in Stampede: 3-6 months instead of 12-16 weeks. The popular SCRUM approach was used as it emphasizes collaboration, functioning software, team self management and flexibility to adapt to business realities (Scrum Alliance, 2016).

The global methodology is composed of three successive phases detailed hereafter:

1. *Big Data Context and Awareness*. In this introductory phase, one or more meetings are organised with the target organisation. A general introduction is given on Big Data concepts, the available platform, a few representative applications in different domains (possibly with already a focus on the organisation domain), the main challenges and main steps. The maturity of the client and a few risk factors can be checked (e.g. management support, internal expertise, business motivation).
2. *Business and Use Case Understanding*. This is also the first phase of CRISP-DM. Its goals are to collect the business needs/problems that must be addressed using Big Data and also to identify one or some business use cases generating the most value out of the collected data.
   This phase is organised based on one or a few workshops, involving the Business Manager, Data Analyst, IT architect and optionally selected specialists, such as the IT security manager if there are specific security/privacy issues that need to be checked at this early stage. Both the as-is and to-be are considered. Specific tools to support the efficient organisation of those workshops are described in Section 4. At the end of this step, an project planning is also defined.
3. *Pilot Implementation of Service or Product*. In this phase, the following implementation activities are carried out in an Agile way:
   - Data Understanding: analyse data sets to detect interesting subset(s) for reaching the business objective(s) and make sure about data quality.
   - Data Preparation: select the right data and clean/extend/format them as required.
   - Modelling: select specific modelling techniques (e.g. decision-tree or neural networks). The model is then built and tested for its accuracy and generality. Possible modelling assumptions are also checked. Based on the results, the model parameters can be reviewed or other/complementary techniques can be used.
   - Evaluation: assess the degree to which the model meets business objectives, using realistic or even real data.
   - Deployment: transfer the validated solution to

Table 1: Main characteristics of first pilot wave.

#	Domain	Volume	Velocity	Variety	Main challenge
1	Life science	20 Go/analysis 2 To/week	High (needs parallel processing)	Business data and traceability (food, pharmaceutical and cosmetic industry)	Product quality
2	Space	Galileo ground segment maintenance (12 EU sites, 16 remote sites)	Medium	Hight: messages, logs	Predictive maintenance of costly equipment. High level of dependability (99.8%)
3	Health	900 beds on 3 sites	Real-time	Several sources and formats	Reduce morbidity and mortality, guarantee confidentiality
4	IT Maintenance	About 3000 servers	High (databases, events, logs,...)	Real-time	Predictive maintenance, cost optimisation

production environment, make sure user can use it (e.g. right visualization, dashboard) and start monitoring (performance, accuracy).

Our pilots are kept confidential. However Table 1 presents the main features of the first four pilots based on the three first ”V” of Big Data (Mauro et al., 2016).

4 LESSONS LEARNED

In this section, we present some lessons learned and related guidelines that are useful to support the whole process and increase the chances of success.

Defining measurable and progressive objectives. Through the deployment of a Big Data solution, a company expects to gain value out of its data. The way to measure this value should be defined right from the business understanding phase, typically by relying on KPIs (Key Performance Indicators). Those company should already have defined those KPIs and be able to measure them. If this is not the case, they should start improving on this: in other words, Business Intelligence should already be present!

Based on this, different improvement strategies can be identified, discussed and result in the selection of a good business case. In the selection process, the gap with the current situation should also be considered, it is safer to keep a first project with quite modest objectives than risking to fail by trying a too complex project that could bring more value. Once a pilot is successful, further improvement can be planned bringing in more value.

From Reactive to Preventive and then Predictive. A common scenario we met in the analysis of data is the need to better anticipate problems or even detect and address early events that could develop into problems. We shortly report about two case studies.

In IT maintenance, a KPI is the total cost of maintenance. Different strategies can be used: simply *reacting* to problems after occurrence, *preventing* against their occurrence based on simple observation like a disk almost full, try to *predict* problems based on observation of behavioural patterns. The predictive solution is better but it should only be envisioned if the preventive solution is present. When considering patterns, most common problems should be addressed first, e.g. disks filling up when backups are performed at the end of a week or a month.

In the health domain, in order to reduce improve the care quality and reduce costs, standardisation is introduced through clinical pathways that describes concrete treatment workflows for patients having identical diagnoses or therapy. This also enables the systematic collection and analysis of patient data. Specific indicators have been defined to precisely track the quality of care and help predicting possible degradation due patient related events (e.g. bad blood parameter) or organisation related events (e.g. service capacity problem). The RDI (Relative Dose Intensity) is such an indicator used in breast chemotherapies. Its value reflect the adherence to the protocol and deviations below the 85% threshold are strongly correlated to the likelihood of relapse.

Using questionnaires for workshops. Conducting a workshop requires to pay attention to many issues while also focusing the discussion on the most relevant ones. A questionnaire can provide an efficient support both as possible preparation before the workshop and as check-list during the workshop. Table 2 shows a few questions about the data to process.

Using Modelling Notations (not to be confused with the data modelling step) is useful to support business and data understanding. During workshops, a whiteboard can be used to sketch models together with the audience. In our experience, data-flow and workflow models help to understand which process is generating, altering, storing or retrieving data. UML class diagrams also help to capture the domain structure

On the other hand, use cases should be avoided be-

Table 2: Some workshop questions about data.

- *Q.UD.1* What are the data sources and data types used in your current business processes?
- *Q.UD.2* What tools/applications are used to deal with your current business processes?
- *Q.UD.3* Are your present business processes performing complex processing on data?
- *Q.UD.4* How available is your data? What happens if data is not available?
- *Q.UD.5* Do different users have different access rights on your data?
- *Q.UD.6* Does your data contain sensitive information (e.g. personal or company confidential data)?

cause they only focus on a specific function and cannot provide a good global picture of the problem.

5 CONCLUSIONS

In this paper, we described how we addressed the challenges and risks of deploying a Big Data solution within companies willing to adopt them to support their business development. Based on different methods and experience reports for the literature, we came up with a method fitting our needs and continuing to evolve as we explore more uses cases, while highlighting a number of lessons learned.

When considering the adoption of Big Data analytics in organisations, what is crucial is the process followed to come up with a method that will maximize the chance of success and will fits the needs of each specific organisation.

Moving forward, we plan to consolidate our work based on what we will learn in the next series of project case studies. So far, we have also focused more on the discovery and data understanding phases. We plan to provide more guidance on the project execution phase when enough pilot projects have reached completion or key milestones.

ACKNOWLEDGEMENTS

This research was partly funded by the Walloon Region through the ”PIT Big Data” project (nr 7481). We thank our industrial partners for sharing their cases and contributions to the method assessments.

REFERENCES

Balduino, R. (2007). Introduction to OpenUP. https://www.eclipse.org/epf/general/OpenUP.pdf.

Bedos, T. (2015). 5 key things to make big data analytics work in any business. http://www.cio.com.au.

Chen, H.-M., Kazman, R., and Haziyev, S. (2016). Agile big data analytics development: An architecture-centric approach. In *Proc. HICSS'16, Hawaii, USA*.

Corea, F. (2016). *Big Data Analytics: A Management Perspective*. Springer Publishing Company, Inc.

Crowston, K. (2010). A capability maturity model for scientific data management.

do Nascimento, G. S. and de Oliveira, A. A. (2012). *An Agile Knowledge Discovery in Databases Software Process*. Springer Berlin Heidelberg.

Frankov, P., Drahoov, M., and Balco, P. (2016). Agile project management approach and its use in big data management. *Procedia Computer Science*, 83.

Gao, J., Koronios, A., and Selle, S. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects. In *AMCIS*.

Gartner (2016). Investment in big data is up but fewer organizations plan to invest. http://www.gartner.com.

IBM (2013). Stampede. http://www.ibmbigdatahub.com/tag/1252.

Kelly, J. and Kaskade, J. (2013). CIOs & Big Data: What Your IT Team Wants You to Know. http://blog.infochimps.com/2013/01/24/cios-big-data.

Lau, L. et al. (2014). Requirements for big data analytics supporting decision making: A sensemaking perspective. In *Mastering data-intensive collaboration and decision making*. Springer Science & Business Media.

Mariscal, G. et al. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Eng. Review*, 25(2):137–166.

Mauro, A. D., Greco, M., and Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, 65(3):122–135.

Nott, C. (2014). Big Data & Analytics Maturity Model. http://www.ibmbigdatahub.com/blog/big-data-analytics-maturity-model.

Rot, E. (2015). How Much Data Will You Have in 3 Years? http://www.sisense.com/blog/much-data-will-3-years.

Saltz, J. and Shamshurin, I. (2016). Big Data Team Process Methodologies: A Literature Review and the Identification of Key Factors for a Projects Success. In *Proc. IEEE International Conference on Big Data*.

Saltz, J. S. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In *IEEE Int. Conf. on Big Data, Big Data 2015, Santa Clara, CA, USA*.

Scrum Alliance (2016). What is scrum? an agile framework for completing complex projects. https://www.scrumalliance.org/why-scrum.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4).